University of Dundee

University of Dundee

DOCTOR OF PHILOSOPHY

Analysis of Colorectal Polyps in Optical Projection Tomography

Li, Wenqi

Award date:
2015

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. May. 2023

Analysis of Colorectal Polyps in Optical Projection Tomography

Wenqi Li

School of Computing

University of Dundee

This dissertation is submitted for the degree of

Doctor of Philosophy

2015

Abstract

Optical projection tomography enables 3-D imaging of colorectal polyps at resolutions of $5 - 10\,\mu$m. This thesis presents image analysis methods for the polyp diagnosis from such images. Specifically, we investigate 3-D texture-based recognition methods, as well as weakly supervised classification methods, for the diagnostic task of discriminating levels of dysplastic change.

Firstly, we build a patch-based recognition system and evaluate both multi-class classification and ordinal regression formulations. 3-D texture representations computed with a hand-crafted feature extractor, random projection, and unsupervised image filter learning are compared using a bag-of-words framework.

Secondly, two novel classification methods are proposed to learn from partially and weakly annotated images respectively. For the partially annotated images, we developed a relevance ranking method to infer the overall classification model using unlabelled contextual image patches. For the weakly annotated images labelled at the image level, we proposed a boosting with regularised tree algorithm to learn the region classifier.

Results on a database of 90 polyps demonstrate that randomly projected features are effective. Discrimination was improved by carefully manipulating various important aspects of the system, including class balancing, output calibration and approximation of non-linear kernels. For the cancer region classification measured by the area under the ROC curve, 0.81 was achieved by training with image level labels, 0.85 by training with eight mouse click annotations per image. They both outperformed the competing methods. 0.88 was achieved by training with the delineated regions.

Acknowledgements

I would like to express my deepest gratitude to my supervisors — Dr. Jianguo Zhang and Prof. Stephen McKenna — for their encouragement, guidance, and unfaltering support throughout this research.

I would like to thank our collaborators from Ninewells hospital — Dr. Maria Coats and Prof. Frank Carey — for providing image data and valuable suggestions.

I am grateful to the computer vision and image processing (CVIP) group at the University of Dundee, as well as Dr. Wei-shi Zheng from Sun Yat-sen University, for fruitful discussions and useful advice.

I am also grateful to the members of my thesis examination committee — Dr. Lewis Griffin, Prof. Janet Hughes, and Prof. Emanuele Trucco — for their helpful suggestions.

Finally, I would like to thank the financial support from the overseas research students award scheme.

Declaration

Candidate's Declaration

I declare that I am the author of this thesis; that, unless otherwise stated, I have consulted all references cited; that the work of which the thesis is a record has been done by the author; and that it has not previously been accepted for a degree; where appropriate, I have acknowledged the nature and extent of work carried out in collaboration with others included in the thesis.

Wenqi Li Liwenqi

2.7.2015

Supervisor's Declaration

I hereby declare that I am the supervisor of the candidate and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Dr. Jianguo Zhang

Jianguo Zhang

02/07/2015.

to my parents

* * *

献给我的父母

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the invention of X-ray technology by Wilhelm Röntgen in 1895, quite a number of imaging methods have been developed as tools for medical purposes, such as X-ray computed tomography, magnetic resonance imaging, and ultrasound. These tools are now routinely used in clinical medicine and research. Analysis of medical images is essential in modern medicine. In recent decades, the amount of medical image data is growing rapidly due to the popularity of the imaging technologies. Automated image analysis tools are desirable in order to help clinicians and researchers in providing accurate and efficient assessments.

This thesis investigates automated methods for histological analysis of colorectal polyps in optical projection tomography (OPT). To our best knowledge, it is the first time the ability of automatic methods based on pattern recognition techniques is explored for OPT images of human tissue. Figure 1.1 illustrates the workflow of OPT polyp image analysis. The colorectal polyps taken during colonoscopy are processed and scanned using the OPT method. The polyp images obtained using OPT, together with manual annotations provided by the pathologists, are studied in this research. The scope of this thesis is indicated by the dashed box.

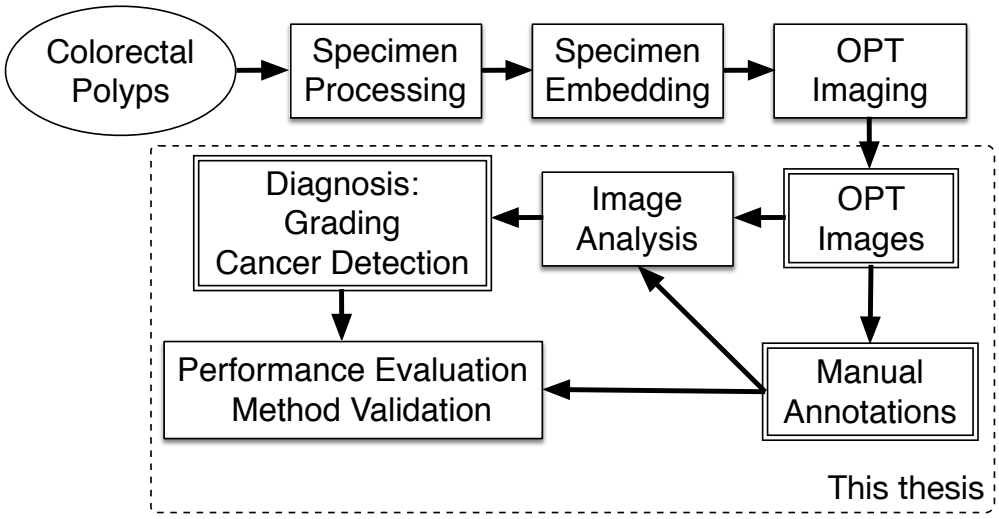
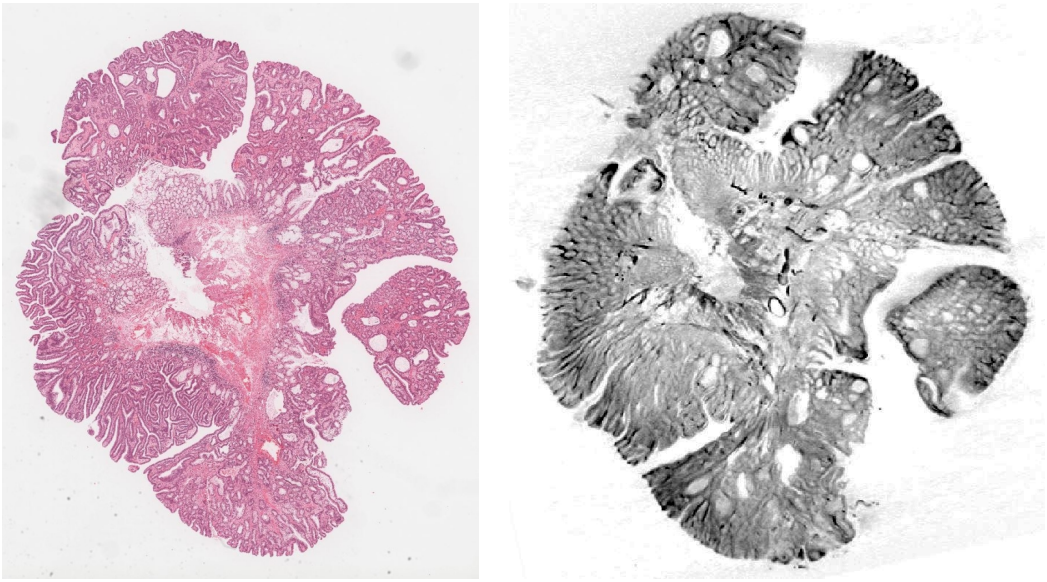


Figure 1.1 A diagram of colorectal polyp imaging and processing in optical projection tomography. The contributions of this thesis lie in the dashed box.

1.1 Motivations

We focus on the task of providing histological diagnosis automatically using colorectal polyp images. Currently, this task is routinely completed by highly trained pathologists who inspect sections of polyp samples stained with haematoxylin and eosin (H&E) under a light microscope. However, this conventional technique has its limitations: (1) it involves taking a thin section of tissue from the centre of the polyp and this will not necessarily be representative of the whole specimen; (2) much variability exists among experienced pathologists when making diagnoses using H&E sections due to features such as epithelial displacement (EPD), in which surface epithelial cells become misplaced into the stalk of the polyp mimicking true invasive cancer. Over-diagnosis of EPD as cancer has a confounding effect subjecting patients to unnecessary treatments and generating inaccurate epidemiology reports [96, 109].

OPT is a relatively new 3-D imaging technique first applied to better our understanding of embryo development [128]. It is a simple and affordable imaging technology that is well-suited for specimens between 0.5 and 10 mm in size. OPT imaging of colorectal polyp is non-destructive to the original tissue and enables virtual sectioning



(a) (b)

Figure 1.2 (a) An H&E-stained whole mount slide. (b) A virtual section from an OPT image of the same polyp with the section chosen to be close to the cut surface that resulted from removal of the physical section shown in (a). Image contrast was manually adjusted for polyp visualisation purpose.

of the specimen at any orientation. Figure 1.2 shows a comparison between an H&E section and an OPT section of the same polyp. A significant advantage of tomography is the flexibility in viewing virtual sections and manipulating the image to gain more information. By contrast, once the cutting angle has been chosen for the H&E section, it cannot be changed. Histology sections are cut from the tissue once it has been embedded in paraffin wax and subsequently stained. These can be viewed at sub-micron resolutions whereas OPT provides a lower spatial resolution of about 5-10 μm [129]. Near-visible light is used to obtain OPT images and therefore polyps must be optically cleared in advance of scanning using benzyl alcohol benzyl benzoate (BABB).

Manual polyp image analysis can be tedious and time-consuming. Additionally, inter- and intra-observer variation exists when pathologists diagnose colorectal polyps, notably when grading dysplasia whether from H&E or OPT images [27, 28]. Reliable and repeatable automatic recognition systems are desirable.

1.2 Research problems

Computer-aided diagnosis systems for colorectal polyps mostly follow a feature extraction and classification paradigm. Analogously, when designing novel recognition systems for optical tomographic images of polyps, two essential research problems arise from a computer vision perspective.

- What 3-D visual feature descriptors are appropriate to represent the histological patterns?
- What discriminative methods are feasible to model the relationship of the features and the diagnostic outputs used by pathologists?

These problems are empirically explored in this thesis, with novel applications of computer vision and pattern recognition techniques.

In the empirical studies of OPT image analysis, large amounts of high-quality manual annotations are needed in order to train an accurate classification model. However, the manual annotations should be generated by experienced pathologists. The process of obtaining annotations can be very costly and laborious. Therefore, a research problem arises in training methods for OPT analysis:

- How to train good-quality histology analysis models with reduced requirements of costly experts' manual annotation?

Two novel machine learning algorithms are proposed in this thesis to train OPT patch classification models. One is designed for efficient training with a form of partial annotation, and the other for both patch and image classification with only image-level annotations.

The aim of this research is twofold. Firstly, to investigate texture-based analysis methods to discriminate diagnostic levels of dysplastic change in OPT colorectal polyp images. Secondly, to propose novel weakly-supervised algorithms for training OPT image classification models. Figure 1.3 illustrates the tasks and the annotations studied

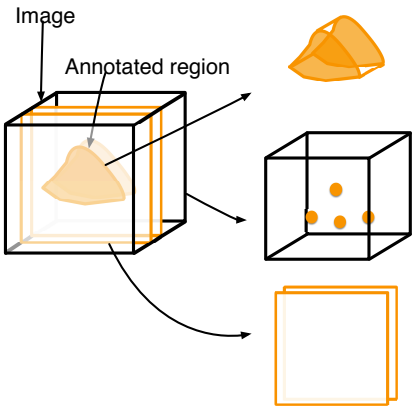
OPT Dataset	Annotation	Task	Chapter
		Texture analysis with region annotations Low-grade dysplasia, High-grade dysplasia, Invasive cancer	Chapter 5 & 6
		Cancer detection with click annotations Invasive cancer vs. non-cancer	Chapter 7
		Cancer detection with image-level annotations Invasive cancer vs. non-cancer	Chapter 8

Figure 1.3 The OPT image analysis tasks and the associated annotations studied in this research.

in this research. Manually delineated regions were used for the patch based 3-D texture analysis. Both mouse click annotations and image-level annotations — as forms of simplified manual annotations — were used to train cancer detectors in OPT images.

1.3 Contributions

The main contributions of this thesis are as follows.

1. It provides the first study in the literature on automatically discriminating between invasive cancer, high-grade dysplasia, and low-grade dysplasia in optical projection tomography images.

2. A rigorous comparative evaluation of three state-of-the-art 3-D texture feature representations was conducted on a large dataset. It demonstrated that random projection performs better than local binary patterns (LBP), a recent, popular hand-crafted feature, and independent subspace analysis (ISA), an automatic image filter learning technique, in terms of discriminating OPT image patches.

3. While the task of discriminating diagnostic levels of dysplastic change can be cast as a three-class classification problem, this ignores the ordinal structure of

these labels. Hence, a *classification* model and an *ordinal regression* model, both based on margin maximisation, are compared and contrasted for this task. These raise issues of class imbalance and output calibration which are explored empirically. Two state-of-the-art strategies for fast approximation of non-linear kernels are also evaluated. Although the focus is on OPT images of colorectal polyps, the analysis and evaluation methods used should be applicable to other ordinal regression tasks in other image modalities.

4. To reduce the requirement of costly manual annotations in training classifiers, we investigated the use of partial annotations consisting of just one or a few clicks in the polyp region of interest. A learning framework using partially annotated OPT images was proposed for colorectal cancer detection. The proposed ranking algorithm enables efficient training of classification models by utilising the contextual information near the clicks.

5. A novel multiple instance learning algorithm was proposed for cancer detection in OPT images using image level annotations, i.e., a binary label indicating whether cancer is present in the image. With images annotated at image-level, we first search a set of region-level prototypes by solving a submodule set cover problem. Regularised regression trees are then constructed and combined on the set of prototypes using a multiple instance boosting framework. This algorithm explored training classification models without using the cancer location information, which brings a further reduction of annotation workload compared to the partial annotations.

1.4 Thesis outline

Chapter 2 introduces optical projection tomography, colorectal polyps and cancer, relevant clinical background, and summarises the image dataset.

Chapter 3 reviews applications of colorectal polyp image analysis in the literature.

Chapter 4 reviews 3-D texture feature extraction in the literature, and describes the three representative texture feature extraction methods we explored.

Chapter 5 presents a novel application of texture analysis methods for discriminating dysplastic regions in OPT colorectal polyp images.

Chapter 6 presents a novel framework using partial annotations for cancer and non-cancer classification in OPT images.

Chapter 7 presents a novel algorithm using multiple instance learning framework, to train cancer detectors with image-level annotations.

Chapter 8 concludes the thesis with discussions of the proposed methods and gives potential directions for future research.

Chapter 2

Background and Dataset

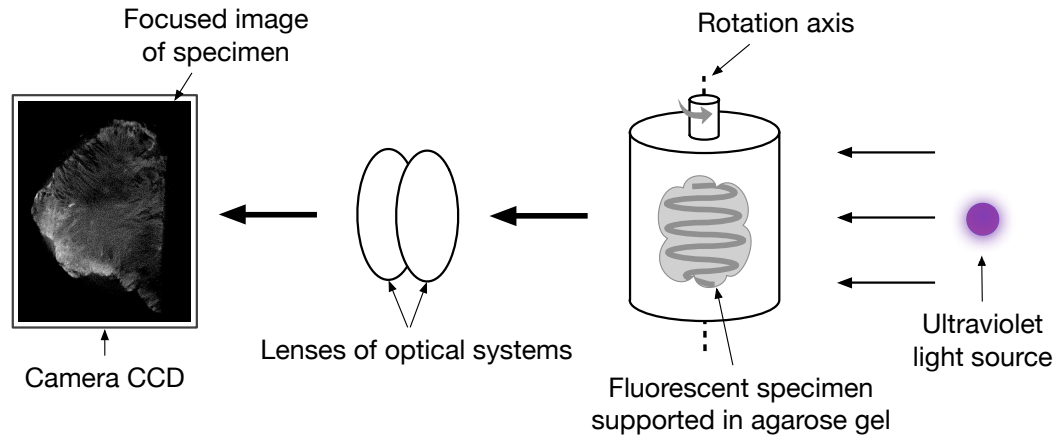
2.1 About this chapter

This chapter introduces the clinical background of the research, including optical projection tomography and its application to colorectal polyp visualisation, colorectal polyp categorisation and colorectal cancer. The data used in this research is also introduced.

2.2 Optical projection tomography

Optical projection tomography is a 3-D imaging method that is ideally suited for specimens between 0.5 and 10 mm in size [129]. OPT can operate in a transmission or an emission mode. The transmission mode involves shining light into the specimen from one side and detecting photons on the other side using a charged-coupled device (CCD). In a sense, the setup of transmission OPT is analogous to the widely used X-ray computed tomography. In the emission mode, the light source is distributed in the specimen by applying fluorescent dyes. When the specimen is exposed to an excitation light (e.g., ultraviolet light), fluorescent molecules are excited throughout

the specimen and emitted photons can be detected by CCD [127]. In this thesis, the images of colorectal polyp were obtained with the emission OPT.



(a)

(b)

Figure 2.1 Optical projection tomography imaging system. (a) A simplified overview of OPT imaging in emission mode (adapted from [128]), (b) A commercial OPT scanner (photo courtesy of `Bruker microCT` [21].)

Figure 2.1 summarises the main features of a typical OPT imaging system. The fluorescent specimen is prepared with optical clearing in advance of scanning using benzyl alcohol benzyl benzoate (BABB) and embedded in a transparent cylinder of agarose gel. The cylinder is rotated around an axis with a fixed small angular step. At each rotation step, a 2-D image is recorded by a CCD chip. After the images are captured throughout 360 degrees, virtual sections can be reconstructed using a back-projection algorithm. The time it takes is about 20 minutes for an OPT scan and about 45 minutes for a 3-D reconstruction process of a 1024^3-voxel image, on a single-CPU PC [128].

Table 2.1 Comparison of 3-D microscopy technologies (partly adapted from [26, 127, 129]).

Modality	Resolution	Pros	Cons
OPT	~ 5 to $10\,\mu$m	Simple setup; affordable; can image both fluorescent and non-fluorescent specimen; image depth up to 15 mm	Specimen needs to be prepared with optical clearing techniques
Serial sections	$\sim 0.2\,\mu$m	Very high resolution	Labour-intensive
Confocal microscopy [147]	$\sim 0.1\,\mu$m	Non-invasive; fast; high resolution	High cost; image depth $\sim$ 1 mm; fluorescent specimens only
OCT [71]	~ 1 to $10\,\mu$m	Can observe living tissues; fast; high resolution	Maximal image depth ~ 2 to 3 mm
μMRI [126]	$\sim 25\,\mu$m	Can observe living tissues; high contrast for unstained specimen	Relatively complex; high cost
Micro-CT [48]	$< 10\,\mu$m	Can observe living tissues; fast	Relatively complex for soft tissue; contrast agent needed

Table 2.1 compares OPT with other 3-D microscopy technologies, including serial sections [135], confocal microscopy [147], optical coherence tomography (OCT) [71], microscopic magnetic resonance imaging (μMRI) [126], and X-ray microtomography (Micro-CT) [48]. OPT is relatively simple, non-invasive and can achieve effective resolutions of 5 to 10 μm. The main disadvantage of OPT microscopy, as compared to μMRI, is that high-resolution reconstruction depends on the specimen being transparent and its tissues possessing a homogeneous refractive index [129].

OPT imaging was first developed to study embryo development [129]. Recently it has been applied to image human tissues and study diseases, e.g., breast tumours [86], liver and pancreas tissues [108], and various cancer cells [2]. Coats et al. [28] investigated the feasibility of using OPT to visualise and diagnose colorectal polyps. In this thesis, we study the automation of the diagnosis. The clinical background of colorectal polypoid cancer is briefly reviewed in the following sections.

2.3 Colorectal cancer

Colorectal cancer was the third most common cancer in men (756,000 new cases per annum, 10.0%) and the second in women (614,000 new cases per annum, 9.2%) worldwide in 2013 [47]. The highest incidence was reported in countries of Europe, North America, and Oceania, whereas incidence was lowest in some countries of south and central Asia and Africa [19]. According to Cancer Research UK [22], colorectal cancer was the third most common cancer in both males and females in the UK; 41,581 new cases were registered in 2011. It is also the second most common cause of cancer death; 16,187 people died of colorectal cancer in 2012. The main risk factors of colorectal cancer include age, male sex, family history of colorectal cancer, and inflammatory bowel disease. Although no single risk factor explains most cases, the most associated factor is age. 95% of the cases occur in people over the age of 50, and the incidence rate strongly increases with age [131].

Colorectal polyps that have malignant potential (adenomas) are the most important precursor lesions of colorectal cancer. About 95% of cancers develop from polyps [19]. However, the majority of polyps remain benign; less than 10% of them develop into cancer. The development from benign polyps to cancer can be very slow: the process may take more than ten years. Detecting and removing premalignant or early-stage cancer polyps is an effective way to reduce mortality of colorectal cancer.

Screening and follow-up of high-risk groups of patients can detect colorectal cancer at an early stage and trigger a potential cure. For example, results from epidemiological studies suggest that sigmoidoscopy (a routine test to examine the lining of a sigmoid colon) screening can reduce incidence and mortality rates of distal colorectal cancer by roughly 60% to 80% [32]. In the UK, the world's largest bowel cancer screening (BCS) programme was started by the NHS in 2006. As of 2012, 65,535 polyps had been excised and recorded in the BCS programme database [100].

2.4 Colorectal polyps and diagnosis

To diagnose colorectal cancer, histological analysis of polyps taken during colonoscopy is essential. The current gold standard method is to examine sections of polyp stained with haematoxylin and eosin (H&E) under a light microscope. By inspecting the H&E sections, pathologists analyse histological patterns of the tissue and identify the cell types. Finally, the findings of the pathologists will decide the treatment for the patients.

A review of the tissue preparation process for routine histology analysis can be found in [105]. The major steps of preparing glass slides are summarised as follows.

(1) *Fixation*. The tissue is first immersed into fixative solutions, then dehydrated and infiltrated with paraffin. This step ensures cells and tissue components are preserved.

(2) *Embedding*. The tissue is embedded into a block of hardened paraffin. The orientation of the tissue is important in this step, as in the next step the tissue block will be sliced parallel to the block.

(3) *Sectioning*. Very thin (a few micrometres thickness) slides of the tissue are cut and transferred into a clean glass slide. Too thick sections can make the nuclei over-stained and the differentiation of cellular components very poor.

(4) *Staining*. In terms of the routine H&E stain, the dyes are applied to the tissue so that they will bind cells and cellular components. Haematoxylin mainly has an affinity with the nucleic acids of the cell nucleus, whereas eosin with cytoplasmic components of the cell.

As mentioned in Section 1.1, the conventional 2-D histology technique has several limitations: a thin section of tissue from the centre of the polyp may not be representative of the whole specimen. Features such as epithelial displacement (EPD) can cause differences between experienced pathologists when making diagnoses using H&E sections.

By contrast, the procedure of conducting OPT scanning of tissue is faster: First embed the specimen in an agarose block, then dehydrate it with methanol, and finally clear it using BABB. A significant advantage of tomography is the flexibility in viewing virtual sections and manipulating the image to gain more information. As a comparison, building a reconstruction of 3-D visualisation with a set of H&E stained sections can be a relatively complex task [135]. However, the H&E sections can be viewed at sub-micron resolutions whereas OPT provides a lower spatial resolution of about 5-10 μm [129].

Coats et al. [28] compared the utility of OPT images and standard H&E stained sections with 352 colorectal polyps from a screening population in the UK BCS programme. The results showed that surface morphology was clearly identifiable by OPT and comparable with the one obtained from H&E; low-grade dysplasia (59.7%) was distinguishable from high-grade dysplasia (25.3%) and invasive cancer (15.0%) using OPT but differentiation between the latter two classes was less distinct. Moreover, OPT demonstrated additional features (e.g., surface ulceration, epithelial misplacement, and vasculature patterns) that were not apparent on H&E sections. Coats et al. [27]

further compared the colorectal polyp diagnostic agreement of OPT and conventional methods among specialist pathologists. 59 specimens (39 low-grade dysplasia; 8 high-grade dysplasia; 12 invasive cancer) specimens were reviewed. Inter-observer analysis showed that no pathologist agreed with the glass slides for all specimens when reviewing glass slides alone. Intra-observer analysis of dysplasia diagnoses showed substantial agreement when comparing glass and digital slides (Kappa coefficient 0.68-0.74; specificity 93.6-100%) and fair to moderate agreement between glass slides and OPT images (Kappa coefficient 0.27-0.47; specificity 86.3-97.6%) Introducing OPT created more variation in diagnoses than using H&E slides. The pathologist with the most OPT experience had the best agreement.

The variations among specialist pathologists suggest that repeatable automated analysis systems are desirable in order to achieve accurate diagnoses. In addition, given the large number of polyps obtained from colon cancer screening in programmes such as the UK BCS, there is a strong motivation for automating the polyp analysis.

2.5 Classification of colorectal polyps

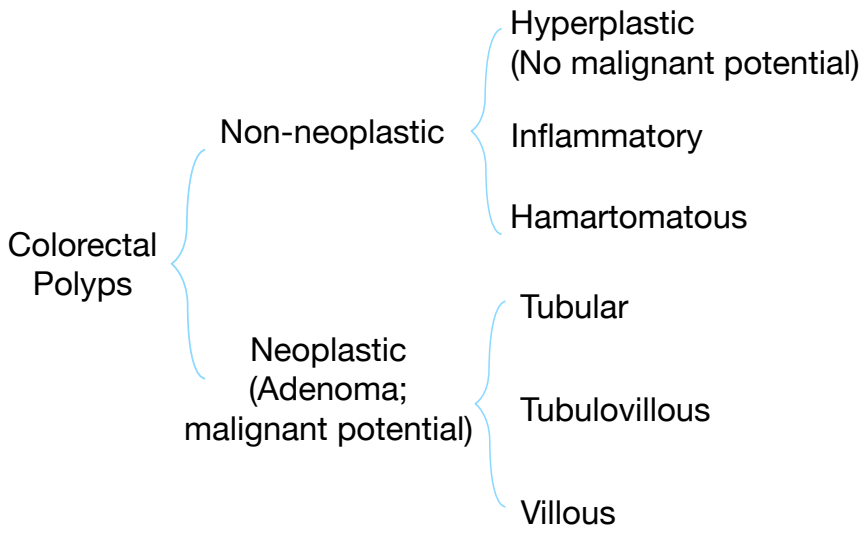


Figure 2.2 A brief classification of colorectal polyps.

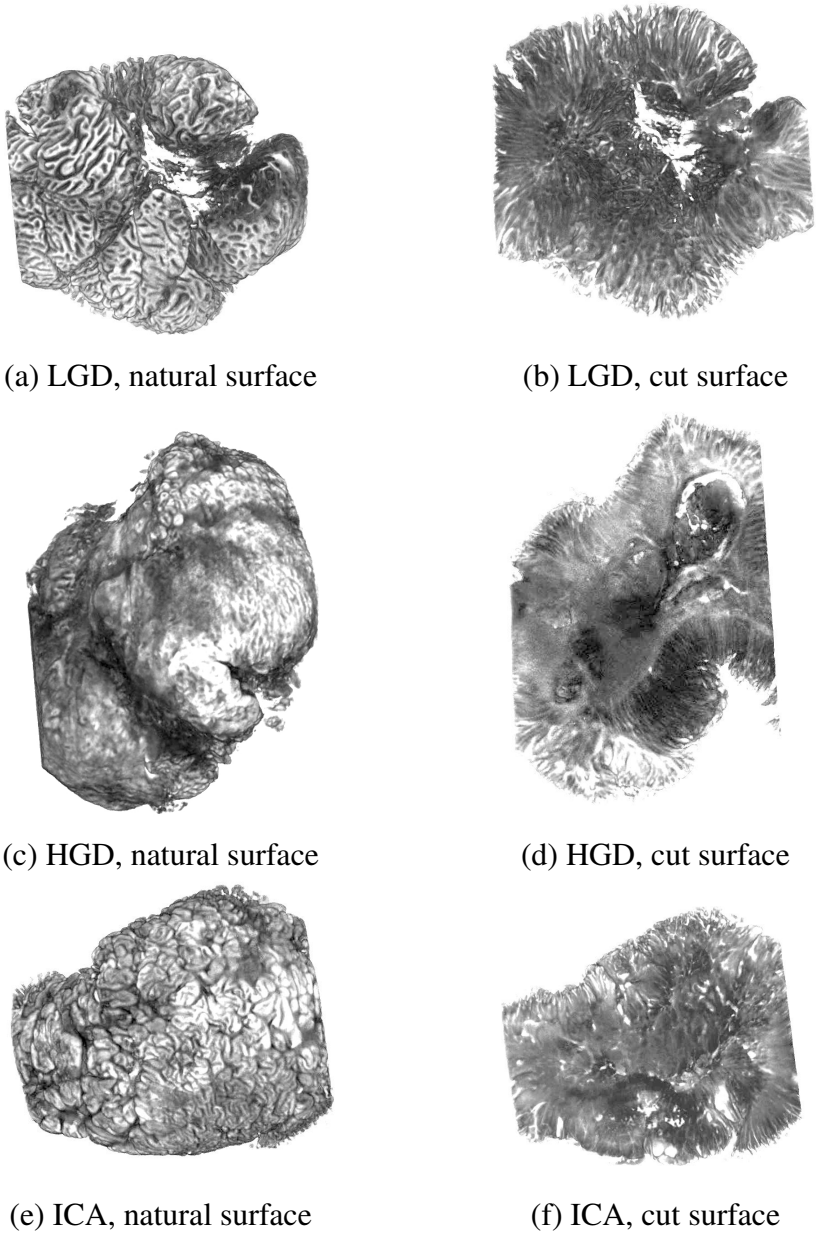
Figure 2.2 briefly summarises the classification of colorectal polyps [31]. Colorectal polyps can be generally categorised into neoplastic and non-neoplastic. Hyperplastic polyps are non-neoplastic usually found in the distal colon and rectum, and are without malignant potential. Neoplastic polyps (adenomas) are benign but have malignant potential. Adenomas can be further divided into three groups based on histological patterns: tubular, tubulovillous and villous adenomas. Tubular adenomas are usually non-advanced neoplastic lesions while the other two subtypes, i.e., tubulovillous adenoma and villous adenoma, are generally more advanced neoplastic polyps.

This thesis mainly concerns adenomas and invasive cancers. Adenoma is characterised by the presence of epithelial dysplasia; it is the most well-known precursor of colorectal cancer. Epithelial dysplasia denotes an unequivocal neoplastic epithelial alteration. It is characterised histologically by: (1) cytological atypia, (2) aberrant differentiation, and (3) disorganised architecture [70]. The grade of dysplasia is an important parameter of adenomas. It measures the aggressiveness of lesions that are considered to be precancerous in terms of microscopic architecture, aberrant differentiation, and cytological features. Measuring the grade of dysplastic change can provide an estimation of the malignant risk that helps preventing and controlling the cancer [46]. Invasive cancer polyps are malignant polyps which invade into the submucosa and beyond, including the metastatic spread.

The term “dysplasia” is only used to describe polyps when there is no evidence of invasion [34]. Following NHS BCS programme lesions reporting guidelines [110], adenomas are graded as *low-grade dysplasia* and *high-grade dysplasia*. Additionally, we include the class of *invasive cancer* polyp where the lesions already show evidence of invasion. The three classes reflect the chronological sequence of dysplastic change and form the output space of our classification system.

2.6 Image acquisition and annotation

2.6.1 Colorectal polyp images



(a) LGD, natural surface

(b) LGD, cut surface

(c) HGD, natural surface

(d) HGD, cut surface

(e) ICA, natural surface

(f) ICA, cut surface

Figure 2.3 Direct renderings of OPT polyp images with polyp voxels rendered as opaque: (a), (b) a low-grade dysplasia (LGD) polyp, (c), (d) a high-grade dysplasia (HGD) polyp, (e), (f) an invasive cancer (ICA) polyp. Top row: viewing angles adjusted to view the natural surfaces of the polyps. Bottom row: viewing angles adjusted to view artefactual surfaces due to physical cuts.

Ninety colorectal polyps were selected from the NHS Tayside Tissue Bank archive to be representative of the subgroups: *invasive cancer* (ICA), *high-grade dysplasia* (HGD) and *low-grade dysplasia* (LGD). Thirty samples were obtained for each of these three groups to give a balanced dataset. The H&E stained sections taken from each specimen were re-diagnosed by an experienced gastro-intestinal histopathologist according to the NHS BCS programme and WHO guidelines to reduce intra-observer bias [66, 110]. Images were acquired using OPT in emission mode under ultraviolet light and Cy3 dye at a voxel resolution of 6.7 μm^3. Each image was of one colorectal polyp and had 1024^3 voxels. Figure 2.3 and Figure 2.5 demonstrate 3-D and 2-D visualisations of tissues from each class in OPT respectively.

2.6.2 Image annotations

Table 2.2 Organisation of the dataset.

Classes	LGD	HGD	ICA
Number of images	30	30	30
Number of annotated slices	2,710	787	2,446

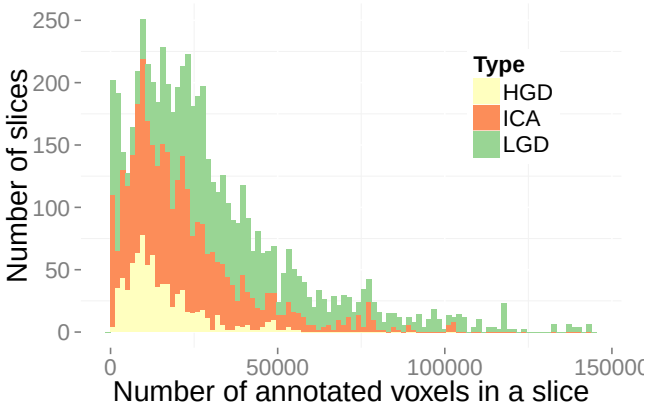
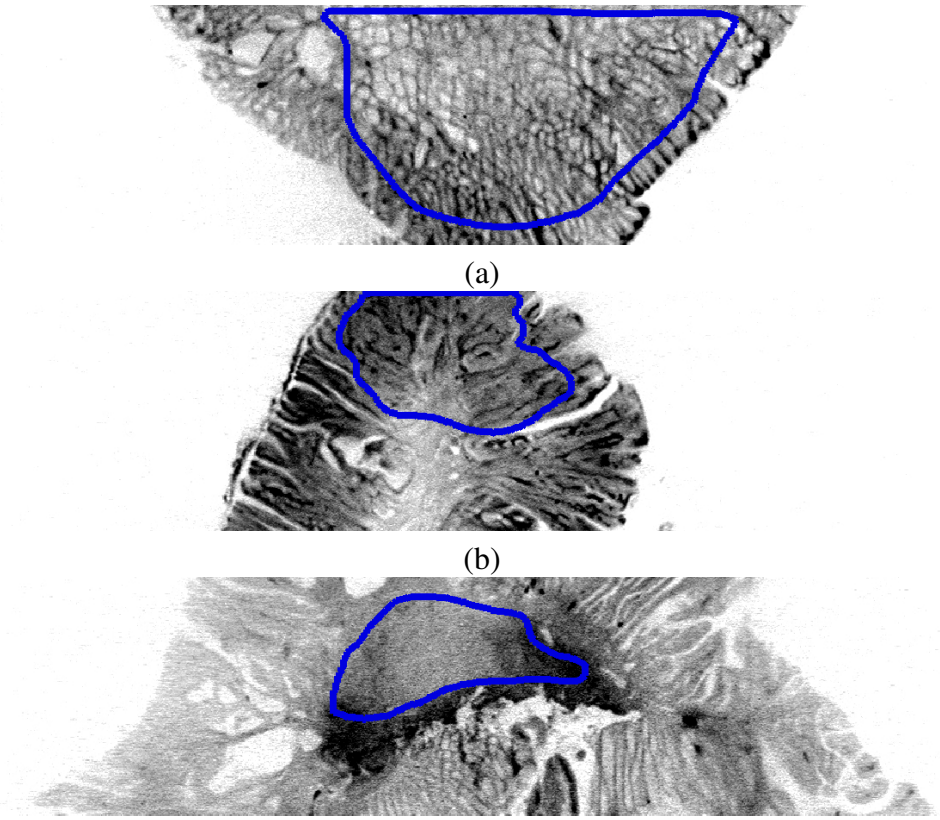


Figure 2.4 Statistics of annotated slices per class.

Each 3-D image was manually annotated with 3-D regions by an individual experienced in interpreting OPT images (Dr. Maria Coats from the University of Dundee's School of Medicine). Characteristic regions were annotated in each polyp, i.e., regions

of ICA were annotated in polyps labelled as ICA, regions of HGD were annotated in polyps labelled as HGD, and regions of LGD were annotated in polyps labelled as LGD. Each region's boundary was delineated such that the annotator had high confidence that all tissues within the region were correctly labelled. The H&E slide corresponding to the cut surface of each polyp was used as guidance for this annotation. Annotations were performed using the software tool ITK-SNAP [153] by delineating 2D regions every 4 or 5 slices and then interpolating between them (as illustrated in Figure 2.6). Table 2.2 and Figure 2.4 summarise the quantities of voxels annotated per slice. Figure 2.5 shows some examples of annotated slices. Although the number of images was the same in each class, the numbers of annotated regions were unbalanced. The issue of training from an unbalanced dataset is addressed in Chapter 5.



(a)

(b)

(c)

Figure 2.5 Images showing slices with regions annotated as (a) LGD, (b) HGD, and (c) ICA. Image contrast was manually adjusted for visualisation purpose.

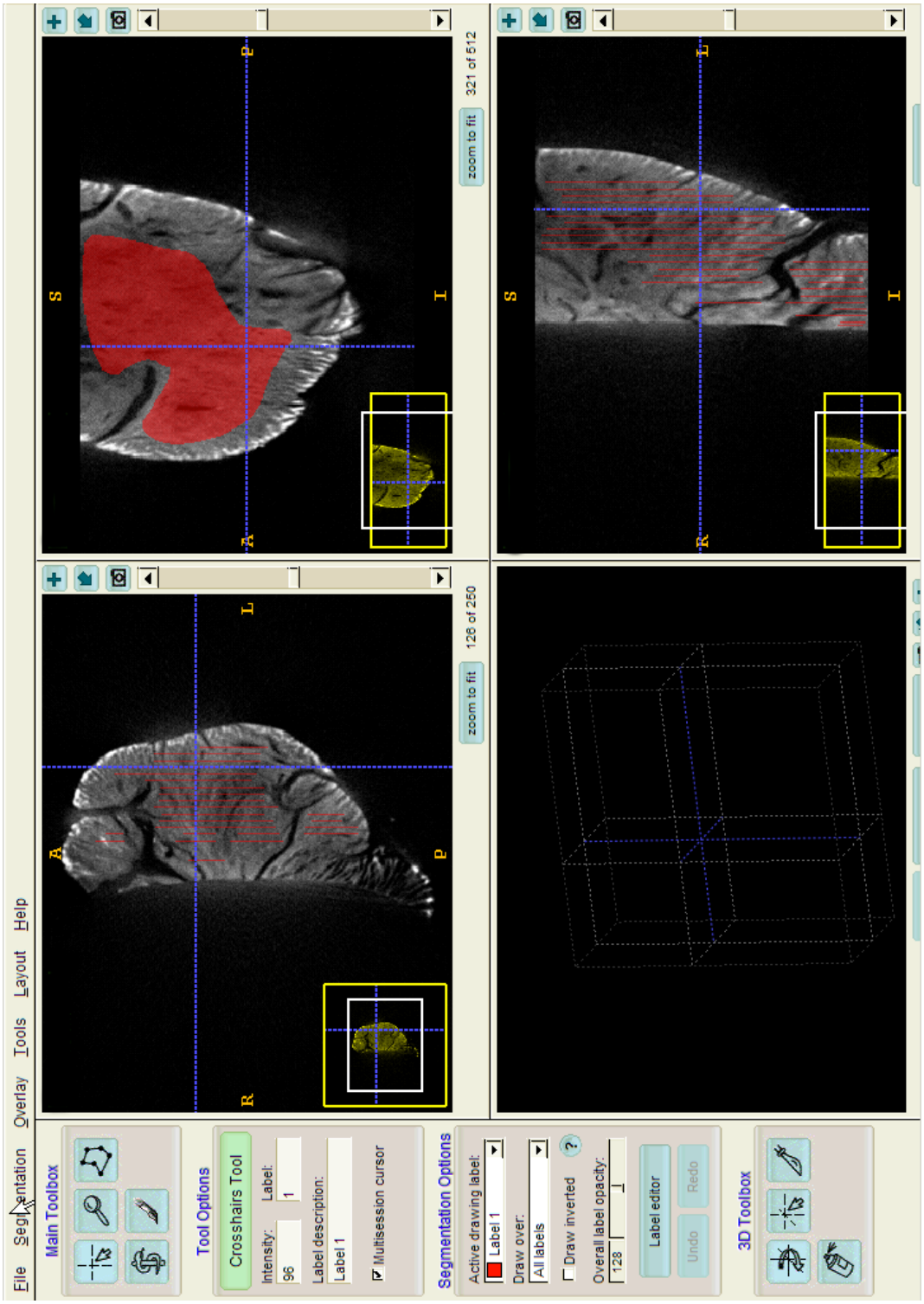


Figure 2.6 Annotating an OPT polyp image using ITK-SNAP [153].

2.7 Summary

This chapter summarises the background of analysing colorectal polyps using OPT method. The advantages of the 3-D imaging method, its potentials for polyp visualisation and cancer diagnosis were emphasised. The images and manual annotations employed throughout this research were introduced. The next chapter reviews the related work of colorectal polyp image analysis.

Chapter 3

Literature review

3.1 About this chapter

This chapter reviews automated methods proposed in the literature for pathological classification of polyps, i.e., for the detected polyps or removed polyp samples, differentiate the pathological stages to which they belong (as illustrated in Figure 2.2). Various imaging methods were demonstrated effective for such purpose. This chapter is organised according to the imaging modalities that were employed in the systems. Table 3.1 lists the colorectal polyp analysis systems according to the type of features and classifiers, as well as imaging modality and dataset information.

Table 3.1 Summary of the related systems for colorectal polyp diagnosis.

Study	Stain	Modality	Pathology	Dataset	Features	Classification
Rajpoot and Rajpoot [117]	H&E	hyperspectral microscopy	normal, cancer	11 images	structure	SVM
Masood and Rajpoot [104]	H&E	hyperspectral microscopy	normal, cancer	32 images	texture	PCA,LDA,SVM
Chaddad et al. [23]	–	hyperspectral microscopy	hyperplastic, neoplastic, cancer	16 images	structure and texture	NN
Kalkan et al. [78]	H&E	microscopy	normal, cancer, adenomatous, inflamed	2,000 patches	structure and texture	logistic regression
Kalkan et al. [77]	H&E	microscopy	normal, cancer, adenomatous, inflamed	120 images	structure and texture	KNN and SVM
Tosun et al. [140]	H&E	microscopy	normal, cancer	150 images	structure and texture	–
Tosun et al. [141]	H&E	microscopy	normal, cancer	16 patients	structure and texture	–
Simsek et al. [132]	H&E	microscopy	normal, cancer	200 images	structure and texture	–

(Continued on next page)

Table 3.1 (continued)

Study	Stain	Modality	Pathology	Dataset	Features	Classification
Altunbay et al. [6]	H&E	microscopy	normal, low-grade, high-grade cancer	213 images	structure	SVM
Ozdemir and Gundus-Demir [114]	H&E	microscopy	normal, low-grade, high-grade cancer	3,236 images	structure and texture	SVM
Ozdemir et al. [115]	H&E	microscopy	normal, low-grade, high-grade cancer	3,236 images	structure and texture	resampling, voting
McCann et al. [106]	H&E	microscopy	normal, inflamed	40 images	structure	NN
Cohen et al. [29]	H&E	microscopy	normal, cancer	200 images	structure	RF
Olgun et al. [112]	H&E	microscopy	normal, low-grade, high-grade cancer	3,236 images	structure	SVM
Rathore et al. [121]	H&E	microscopy	normal, cancer	174 images	structure and texture	RF, boosting
Rathore et al. [119]	H&E	microscopy	normal, cancer	174 images	structure and texture	ensemble of SVM
Rathore et al. [120]	H&E	microscopy	normal, cancer	174 images	structure and texture	SVM
Ahmad et al. [3]	H&E	microscopy	normal, cancer	255 images	structure	NN

(Continued on next page)

Table 3.1 (continued)

Study	Stain	Modality	Pathology	Dataset	Features	Classification
Xu et al. [150]	H&E	microscopy	normal, 4 classes of cancers	103 images	texture	MILBoosting
Gan et al. [52]	H&E	microscopy	normal, adenomatous, cancer	100 images	texture	genetic algorithm
Hamilton et al. [65]	H&E	microscopy	normal, adenomatous	20 images	texture	FS
Esgiar et al. [42]	H&E	microscopy	normal, cancer	21 images	structure and texture	–
Shuttleworth et al. [130]	H&E	microscopy	normal, adenomatous, cancer	175 images	texture	–
Atlamazoglou et al. [13]	fluorescent	microscopy	normal, cancer	70 images	texture	LDA
Esgiar et al. [43]	IHC	microscopy	normal, cancer	102 images	texture	KNN
Song et al. [133]	–	CT colonoscopy	neoplastic, non-neoplastic	110 images	texture	SVM
Tamaki et al. [138]	–	NBI endoscopy	pit pattern	1412 images	texture	SVM
Tischendorf et al. [139]	–	NBI endoscopy	neoplastic, nonneoplastic	209 images	structure	KNN, SVM

(Continued on next page)

Table 3.1 (continued)

Study	Stain	Modality	Pathology	Dataset	Features	Classification
Stehle et al. [137]	–	NBI endoscopy	adenomatous, hyperplastic	56 images	structure	linear classifier
Kwitt and Uhl [82]	dye-spraying	chromo-endoscopy	pit pattern	484 images	texture	NN
Häfner et al. [60]	dye-spraying	chromo-endoscopy	pit pattern	484 images	structure and texture	NN
Häfner et al. [56]	dye-spraying	chromo-endoscopy	pit pattern	627 images	structure and texture	FS, KNN
Häfner et al. [64]	dye-spraying	chromo-endoscopy	pit pattern	627 images	texture	NN
Häfner et al. [63]	dye-spraying	chromo-endoscopy	pit pattern	327 images	texture	KNN
Häfner et al. [62]	dye-spraying	chromo-endoscopy	hyperplastic, adenomatous	716 images	texture	KNN
Fu et al. [50]	–	colonoscopy	hyperplastic, adenomatous	365 images	texture	FS, SVM
Mitrea et al. [107]	–	ultrasound	cancer, inflamed	130 images	texture	RF, SVM, Adaboost
Shao et al. [125]	–	spectroscopy	normal, hyperplastic, adenomatous	198 images	–	PCA, LDA

(Continued on next page)

Table 3.1 (continued)

Study	Stain	Modality	Pathology	Dataset	Features	Classification
Rodriguez-Diaz et al. [123]	–	spectroscopy	normal, hyperplastic, adenomatous, cancer	494 images	–	PCA, SVM
Ayoub et al. [14]	–	stereo endoscopy	normal, cancer	111 images	texture	SVM
André et al. [10]	–	endomicroscopy	benign, neoplastic	1036 images	texture	KNN
André et al. [8]	–	endomicroscopy	benign, hyperplastic, tubular, tubulovillous, cancer	121 videos	texture	KNN
Zhang et al. [157]	fluorescent	OPT	pit pattern	28 images	structure and texture	SVM, RF, KNN, naive Bayes

"–" indicates no information available or not applicable. "SVM" indicates support vector machine; "NN" indicates neural network; "KNN" indicates *k*-nearest neighbour classifier; "RF" indicates random forest; "PCA" indicates principal component analysis; "FS" indicates feature selection; "LDA" indicates linear discriminant analysis.

3.2 Imaging modalities for polyp diagnosis

Microscopic imaging is the most widely used tool in histological analysis. By examining the high resolution (sub-micron) visualisations of the polyp sections obtained with microscopy, accurate staging of colorectal cancer can be made. At this resolution, cellular components and subcellular details of colon biopsies are available. Previous work studied modelling texture features as well as structure features of the cells or cellular components in the images. A large amount of histopathology analysis literature exists for colon biopsies [118], and human tissues in general [53]. Here we review related applications in terms of system targets, i.e., providing pathological grades and cancer region detections in colorectal polyps.

At considerably lower resolutions, modalities including routinely used endoscopy and computed tomographic (CT) colonography can be used for histological analysis. Relatively new approaches — such as narrow band imaging (NBI) zoom-videoendoscopy, capsule endoscopy, and microendoscopy — were also explored in the literature. These methods are more convenient and less invasive compared to the 2-D microscopy. However, the resolutions are limited: the cellular level details are hardly identifiable. Based on these modalities, extensive computer-aided diagnosis (CAD) systems were proposed for polyp detection and diagnosis. Liedlgruber and Uhl [93] reviewed endoscopy-based systems; Yoshida et al. [151, 152] reviewed systems using CT colonography.

3.3 Microscopy image analysis

The analysis of microscopy images can be roughly divided into segmentation-based and segmentation-free methods.

Segmentation-based methods

In segmentation-based methods, pixels are first grouped into different regions corresponding to cells or other components. Then the image can be represented by the set of descriptive statistics of each individual segment.

To detect meaningful structures in the images, the most common pre-processing step is to segment images using colour information. For example, image pixels can be clustered according to colour using k-means algorithm [6, 77, 78, 106, 117, 121, 132, 140, 141]. Usually k is set to 3 because white, pink and purple are the three main colours in an H&E section and roughly correspond to the lumen, connective tissue, and nuclei. Similarly Ozdemir et al. [114] and Olgun et al. [112] used colour deconvolution and quantisation to divide each image into 2 or 3 different parts.

After applying the k-means segmentation, Rajpoot and Rajpoot [117] extracted multi-scale morphological features (such as area, eccentricity, equivalent diameter) from the segmented image. In their experiments, the morphological features were classified into normal or malignant using support vector machines (SVMs) with a Gaussian kernel. The morphological features performed better than statistical features (such as patch location, pixel variation measures, and some high-order statistics) mainly because the former were gathered from the segmented tissue cell image. Masood et al. [103] further studied the selection of the best spectral band for the hyperspectral images and also compared texture features and morphological features. The texture features they applied is the circular local binary patterns (LBP). They showed that using a single spectral band with the texture features can achieve better classification performance than using morphological features. This suggests LBP-based features can efficiently represent the patterns of the histology components. McCann et al. [106] followed the idea of patch encoding using features extracted from the segmented images. They first applied moment filters to detect nuclei from the segmented images; then they developed a set of features based on visual cues used by pathologists, such as nucleus size, colour, density, as histopathology vocabulary to describe image patches. A

pixel-level classifier was constructed using an artificial neural network to classify each pixel and its supporting region. To identify abnormalities at image level, a multiple-instance inference rule was used, i.e., if and only if there exist abnormal patches, the image is labelled as abnormal; otherwise the image is normal. However, these methods did not fully utilise the shape information of cell components in the segmented images. In Kalkan et al. [78] a Laplacian of Gaussian blob detector was used to detect nuclei in H&E-stained sections. A combination of statistics computed from the detected nuclei and texture features were used as the final image patch representation.In their experiments, these image patches were classified as normal, cancer, adenomatous and inflamed classes. The authors applied feature selection and observed that nuclei shape features, Haralick features, and Gabor filter features were the most important features for patch classification. The patch classifiers were further extended to image level by averaging patch classification scores and applying logistic regressions [77]. Rathore et al. [121] represented the segmented images with run-length features and percentage of clusters area features for the purpose of normal and malignant region classifications. They showed that ensemble classifiers such as random forest, rotation forest and boosting classifiers were accurate for their classification tasks.

More recently, capturing structural information by constructing a graph using the segmented H&E image showed promising results [6, 112, 114, 132, 140]. A typical procedure of generating a graph is as follows: first, quantifying image pixels into three groups (e.g., white, pink and purple in H&E sections) using k-means; then, applying a circle-fit algorithm to locate a set of circular objects; finally, a colour graph with three types of colour node can be constructed by applying Delaunay triangulation to the centroids of the circular objects. Using this method, the spatial distribution of the circular objects can be encoded efficiently. Altunbay et al. [6] proposed to quantify a colour graph with three global properties of the graph: the average number of edges for each node, the average clustering coefficient reflecting the connectivity of the nodes, and the diameter quantifying the paths between the graph nodes. Their methods with

colour graph features demonstrated better performance in classifying normal, low-grade, and high-grade images compared to statistical features, and Haralick features. Tosun et al. [140] defined a graph-edge run-length matrix from the colour graph and calculated statistical features such as short run emphasis and gray-level nonuniformity. A region growing algorithm was then employed to segment the H&E sections in an unsupervised manner. Ozdemir et al. [114] proposed to first select a set of subgraphs as query graphs, then use graph edit distance to the query graphs as image features. Simsek et al. [132] measured co-occurrences of the graph nodes and subgraphs at multiple scales for cancer region segmentation. Olgun et al. [112] proposed local object patterns to encode the n-nearest neighbours of the circular objects as features.

Apart from the colour-based segmentation methods, the active contour method was also studied in the literature for tracking the boundaries of the glands and nuclei [3, 23, 29, 30].

In general, the segmentation-based methods require considerable domain knowledge (e.g., shape of nuclei and glandular object, colour of cell components under different stain materials) in order to detect and analyse meaningful image segments. However, compared to the widely used microscopy images, OPT imaging technique is still in its infancy; designing the colour and shape based feature extractors according to the morphology is not feasible in the current study.

Segmentation-free methods

In segmentation-free methods, discriminative features are extracted from the patches or the images directly, without localising cells and other components. To apply texture feature extraction, the images are usually divided into regular grids. The features are extracted from each grid patch and then aggregated into the final patch or image representation. The representation is then classified by general-purpose classifiers [13, 42, 65, 130]. Masood and Rajpoot [104] adopted circular LBP as image patch representation and showed that support vector machines outperformed subspace meth-

ods such as principal component analysis and linear discriminant analysis. Ozdemir et al. [115] augmented their image dataset using resampling-based Markovian model and demonstrated the efficiency of second-order statistics on co-occurrence matrices with the bag-of-words approach. Fractal analysis combined with a nearest neighbour classifier was used by Esgiar et al. [43], who showed that, compared to the conventional texture features (correlation and entropy features), the fractal dimension improves sensitivity and specificity in colon cancer detection. Lim et al. [52] extracted co-occurrence matrix features and classified by a genetic algorithm. Rathore et al. [119, 120] exploited histogram of oriented gradients (HOG), colour component based statistics, and Haralick features for normal and malignant image classification. Ensembles of linear and non-linear SVMs were used as feature classifiers. Xu et al. [150] encoded histopathology image patches with texture features and proposed multiple instance learning algorithms to classify the image patches. The algorithm clustered and classified image patches simultaneously with only image-level annotations. This is different from many systems in the literature that require detailed pixel-level annotations for patch classifications.

3.4 Endoscopic image analysis

Studies showed that the macroscopic surfaces of polyps were also useful for histological diagnosis, e.g., Kudo et al. [81] proposed and later modified [75] the pit pattern classification scheme for polyp classification. For diagnostic purposes colorectal polyps can be divided into six groups, based on the morphological features of mucosal surfaces (pit patterns I-V, illustrated in Figure 3.1). In the literature, a number of systems were proposed to analyse polyp surfaces in vivo following the pit pattern scheme. Texture feature extraction is one of the key ingredients of the systems. The rest of this section briefly reviews the related approaches.

One of the most active research groups in this area is the group of the EndoPit project at the Medical University of Vienna. In their studies, the images were obtained

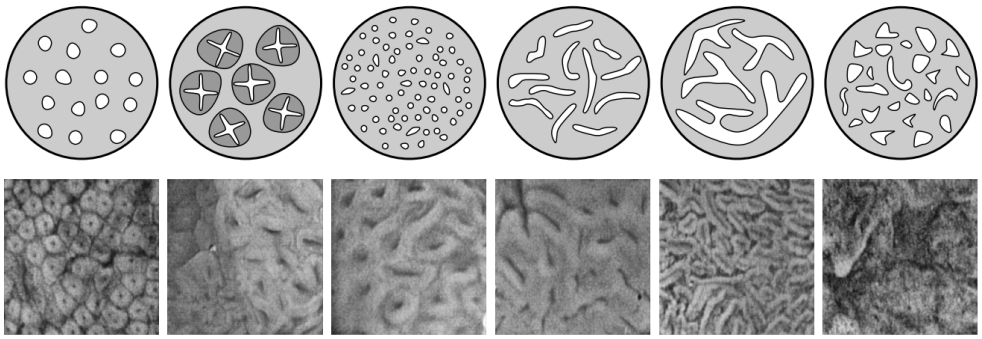


Figure 3.1 Illustration of the pit pattern characteristics (first row) and examples of colorectal polyp images obtained during a high-magnification colonoscopy (second row). From left to right columns, the images correspond to pit pattern I, II, III-S, III-L, IV and V, respectively (courtesy of Häfner et al. [55]).

with a magnifying endoscope and a dye-spraying procedure (chromo-endoscopy). The magnification, with up to 150 times zoom factor, can reveal fine details of mucosal surfaces as well as small lesions. Dye spraying was employed to enhance the visual appearance. The dye used in their experiments is indigo-carmine. The methods proposed by the group include (1) frequency domain features extracted using Fourier transform or wavelet transform, (2) extensions of LBP, and (3) texture features by detecting pit candidates [54–63, 82, 83]. The methods are summarised as follows.

Frequency domain features

Häfner et al. [60] and Kwitt et al. [82] studied several types of discrete wavelet transforms for textural feature representation. 2-D wavelet transforms were applied to each image to capture the image details oriented approximately 15°, 45°, 75°, 105°, 135°, and 160°. The magnitudes of the complex wavelet coefficients are approximately shift-invariant. Conventionally, statistics of the coefficients such as the empirical mean and the empirical standard deviation were used as texture features. Kwitt and Uhl [82] proposed to model the marginal distributions of the wavelet coefficients and use the parameters of the probability density function as novel texture descriptors. Häfner et al. [60] extended the co-occurrence matrix based method to the wavelet-domain and further calculate Haralick features from the matrix.

Variants of local binary patterns

Häfner et al. [57] and Häfner et al. [62] argued that the standard LBP is not suitable for noisy endoscopic images. Two extensions of LBP, multi-scale block LBP (MB-LBP) and local colour vector patterns (LCVP) were proposed for endoscopic texture analysis. LBP operator encodes a local image patch as a binary code, by thresholding the central pixel and its neighbours. MB-LBP was proposed to use neighbouring image blocks instead of pixel neighbours which reduce the influence of noise. Choosing different block size changes the scale of the feature. LCVP was proposed for colour images where intensity values used in LBP were replaced by 3-D colour vectors consisting of values of the three colour channels. The pixel intensity value comparisons in LBP were extended to colour vector comparisons. These features were classified using a k-nearest neighbour classifier. The performance of combining LCVP and MB-LBP was better than the standard LBP and LCVP in the classification of polyp surfaces according to the pit pattern scheme.

Edge-based features

Instead of using texture features, edges of pits were detected in [56]. This was achieved by first removing artefacts using anisotropic diffusion, and then applying Canny edge detection. Given the edges, 18 statistics — including number of pits, mean area of all pits, and mean of average intensity within each pit — were calculated as the image representation. Greedy forward feature selection and k-nearest neighbours classifier were jointly applied to classify the images. An overall classification accuracy of 97% for two classes and 88% for six classes was reported for the edge-based method, while with the above-mentioned wavelet-based features the accuracy was 99% and 96% respectively.

Other work from the EndoPit project

Häfner et al. [59] evaluated colour histograms constructed for the three colour channels as well as co-occurrence histograms that count the number of pairs of pixels at given separation distances as the image features. In Häfner et al. [63], MB-LBP operator was applied to detect pits, and a graph was constructed using Delaunay triangulation to represent the spatial distribution of the pits. Finally, the density of the pits was measured by histogramming the edge lengths of the graph. The histogram was the final representation of polyp surfaces.

Recently Häfner et al. [61] conducted experiments to compare the methods mentioned above in a consistent cross-validation scheme. An extension of LBP achieved the highest overall accuracy in their evaluations. Although the dataset in the experiments was relatively small, it suggested that the LBP based method is very promising in encoding the textural appearances of the polyp surfaces.

In terms of the classification methods, Häfner et al. [58] compared k-nearest neighbour (KNN) and support vector machine (SVM) classifiers for the pit pattern classification using colour histograms. KNN showed better classification accuracies than SVM in their experiments. Häfner et al. [55] further considered solving the multi-class pit pattern classification problem in a one-vs-one classification scheme with a nearest-neighbour classifier and optimising each classifier using a greedy feature selection. The authors argued that by using one-vs-one classification scheme the multi-class problem is decomposed into simpler and easier binary problems. The results showed a remarkable improvement compared to KNN. More recently, Kwitt et al. [83] modelled each image as a collection of local features which were sampled from a set of pit pattern concepts. Gaussian mixture models were used to estimate the distribution of each concept. Each image was map onto a semantic space of pit pattern concepts and classified with a kernel SVM. They argued that it is possible to learn a set of semantically meaningful visual concepts corresponding to the pit pattern scheme. However, it was not clear how to decide the number of concepts in their algorithm.

Apart from analysing the chromo-endoscopy images, Tamaki et al. [138] proposed to analyse polyp surfaces with images obtained by NBI zoom-videoendoscopy. Compared with chromo-endoscopy, NBI image is more convenient as it does not require spraying, washing, and vacuuming dye and water. Tamaki et al. [138] proposed to encode NBI images with densely extracted local features and then calculated histograms using the bag-of-words framework. Specifically, SIFT descriptors were extracted on a regular grid over the image; then the authors proposed to calculate the difference of SIFT descriptors at adjacent grid points as local features (gridSIFT). SVM was used to map the bag-of-words of gridSIFT features to five class labels in the pit pattern scheme. They showed that the bag-of-words histogram of gridSIFT with linear SVM is sufficient for polyp classification.

Similarly, with a bag-of-words framework, André et al. [8–10] investigated images obtained with probe-based confocal laser endomicroscopy, which is able to visualise the epithelium at microscopic level during the endoscopy procedure. In their endomicroscopic database, an image of diameter 500 pixels corresponds to a field of view of 240 μm. To represent benign and neoplastic images, the authors proposed to extract dense SIFT features from the disc regions and the regions were further encoded with the bag-of-words framework. k-nearest neighbour classifier with χ^2 similarity measure was used to classify the bag-of-words histogram representations. In their experiments, a binary classification was conducted to differentiate neoplastic and non-neoplastic while a multi-class classification was conducted to classify five types of lesions: benign, hyperplastic, tubular, tubulovillous, and cancer. They demonstrated the proposed densely extracted features outperformed textons, Haralick features and sparse SIFT features, in terms of classification accuracy. These work has demonstrated that using the bag-of-words encoding method to summarise a set of densely sampled local features is very effective for the colorectal polyp pit pattern classification.

Instead of using the pit pattern scheme, Tischendorf et al. [139] and Stehle [137] explored the feasibility of polyp diagnosis based on segmentation-based vascular

features extracted from NBI endoscopic images. The classification accuracy was 87% on a dataset of 209 polyps (49 non-neoplastic, 160 neoplastic); the result was lower than the human investigators (accuracy 91%).

3.5 Analysis with other imaging methods

Except the widely studied microscopy and endoscopic image analysis, several other imaging modalities were applied to colorectal polyp analysis in the literature. The statistical and textural feature analysis methods are the most common tools.

Song et al. [133] studied 3-D intensity-based textural features for the classification of colon lesions in CT colonography. The Haralick texture model was originally designed for 2-D grayscale images. In [133], the co-occurrence matrix was expanded to 3-D, and 13 statistics were considered in $3\times3\times3$ cubic neighbours. The authors argued that even though the microscopic pathological patterns may not be exactly reflected by the voxel-level patterns in the macro-level images, certain tissue-level texture information may be embedded in the voxel-level pattern. The pattern distributional information can be captured and analysed using the voxel-level feature extraction methods.

A 3-D active sensor was applied to created real-time 3-D reconstruction of colorectal polyps by Ayoub et al. [14, 15]. The polyp was visualised as a cloud of 3-D points using the stereo sensor signals. Ayoub et al. [14, 15] calculated statistics including mean, variance and skewness as features and classified them into hyperplastic and adenomatous with SVMs.

Shao et al. [125] analysed adenomas, hyperplastic polyps as well as normal tissues with images of near-infrared autofluorescence spectroscopy. Near-infrared light can have tissue penetrations up to 1 mm. The fluorescence spectral space consisted of 325 intensity values. Principal component analysis and linear discriminant analysis were applied to classify the returned signals. The classification accuracy on a set of

116 observations was 88.9%, 85.4% and 91.4% for normal tissue, hyperplastic and adenomatous polyp respectively. Similarly in Rodriguez-Diaz et al. [123], principal component analysis was used to reduce the dimensionality of features. An ensemble of linear SVMs was constructed with each trained on a region of the spectrum. The outputs of SVMs were then combined with majority voting and naive Bayes method.

Mitrea et al. [107] applied texture analysis approaches to classify colorectal tumours and inflammatory bowel diseases using ultrasound imaging. They constructed co-occurrence matrix and extracted Haralick features. In the classification process, SVMs, decision trees, random forest and boosting methods were compared; the boosted decision trees achieved the best accuracy among the other classifiers.

3.6 Summary

Most systems in the literature followed the feature extraction and feature classification paradigm. Figure 3.2 compares the colorectal polyp analysis systems according to Table 3.1. Both structure and texture features were widely applied and showed promising results across different imaging methods. For structure features, a detection or segmentation step is usually required in order to identify structures, e.g., cellular components in microscopic images, pits or vessels on polyp surfaces. The classification performance can be largely dependent on the segmentation performance. In comparison, texture features are relatively simple and flexible. Additionally, encoding the densely sampled features with the bag-of-words framework has shown very promising results in polyp classifications in recent work [8–10, 138].

Zhang et al. [157] demonstrated that polyp surface analysis with OPT using structure-based features according to the pit pattern scheme was feasible. This thesis focuses on histological grading using volumetric texture information in OPT images. Only limited work such as [14, 15, 133] in the literature studied texture features in the three-dimensional imaging space for polyp diagnosis. In these texture-based methods,

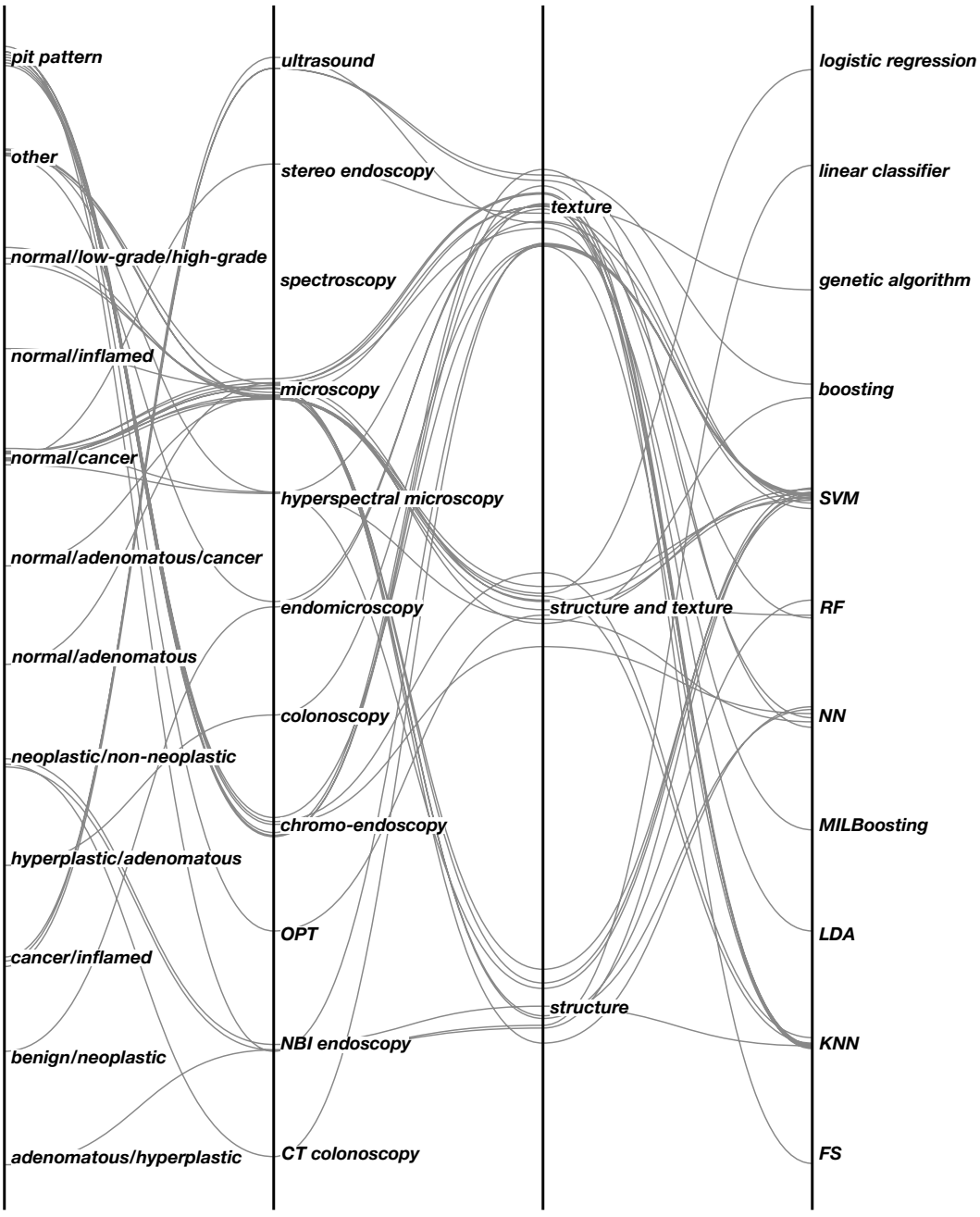


Figure 3.2 A parallel coordinate plot of Table 3.1. The columns from left to right correspond to pathology, modality, features, and classifications, respectively.

relatively simple features such as colour histograms and statistical features (mean, variance, and skewness) were employed.

In terms of classification methods, SVM and KNN classifier were the most popular choices. SVM often showed favourable classification performance, especially in binary classification problems; KNN classifier is relatively easy to implement, and inherently suitable for multi-class classification problems.

Annotations from human investigators were very important in order to train accurate classification models. For relatively new imaging methods, because of lack of established classification standard, histopathology analysis was usually conducted in order to generate ground truth. Recently Xu et al. [150] proposed a weakly supervised learning framework to reduce the annotation requirements in training a microscopic image patch classifier.

In this research, 3-D OPT textures are analysed with 3-D local binary patterns, random projection, and independent subspace analysis techniques using the bag-of-words encoding. The next two chapters (Chapter 4 and Chapter 5) present the details of these methods and also empirical studies in both classification and ordinal regression formulations.

Chapter 4

Feature extraction from optical tomographic images

4.1 About this chapter

In this chapter, for the purpose of discriminating polyp dysplasia in OPT, three representative textural features — 3-D local binary patterns (LBP) descriptors, randomly projected features, and independent subspace analysis features — are experimentally compared. These are state-of-the-art methods from three important categories: hand-crafted feature extractor, randomly generated image filters, and feature representations learned automatically. These methods were selected from the vast literature on texture features mainly because they are among the most advanced and arguably effective for OPT classification. Computational efficiency was also considered when choosing these methods since the OPT datasets in our experiments are large.

Random projection was reported effective in representing general 2-D textures [94]. It can be extended to 3-D and computed efficiently. Local binary patterns are invariant to local contrast changes and can be made invariant to local rotations; these properties are potentially capable of representing micro-textures in OPT images adequately. An approximation of 3-D LBP was employed in our experiments which is not fully invariant

to arbitrary 3-D rotations but computationally feasible. Independent subspace analysis can be used to learn a set of phase- and shift-invariant filters that have similarities to features computed by complex cells in the V1 area of primate visual cortex. ISA learning scales well to large training sets, e.g., [85]. The next section briefly reviews related work on texture analysis. The following sections present technical details of the feature extraction procedures.

4.2 Related work

Previous work demonstrates that texture-based analysis is often an important component of lesion detection, segmentation and classification. 3-D texture features have been widely used in medical image analysis broadly; a comprehensive review of 3-D texture analysis methods is available elsewhere [37].

LBPs are popular, computationally simple texture descriptors that summarise micro-patterns in images. An LBP feature extractor is also an important component of many successful 3-D medical image classification systems under different imaging modalities, e.g., for brain white matter lesion classification in MRI [113], analysis of lung CT [136], fluorescent cell image classification [101], and retinal optical coherence tomography [95].

Random projection (RP), as a non-adaptive dimensionality reduction tool from the compressive sensing theory, was recently applied to image analysis. The motivation of using RP in texture classification comes from the work by Varma and Zisserman [143] in which they demonstrated that using raw patches outperforms using filter banks in the bag-of-words framework. RP is proven to reduce the dimensionality of a raw patch while guaranteeing bounded distortion in k-means clustering analysis [17]. In [17] RP as a simple yet efficient method demonstrated promising results in image clustering tasks. RP's performance was comparable to more sophisticated dimensionality reduction methods including PCA and local linear embedding on a clustering task of a face

images collection. In the classification experiments by Liu et al. [94], RP was applied to 2D texture analysis in a bag-of-words framework. The computationally simple method outperformed the state-of-the-art. Bingham et al. [16] used RP as a dimensionality reduction tool for image classification tasks. In [4], RP accelerated the feature-based registration process of 3-D neural ultrastructure with electron microscopy.

Image filters learned automatically from images have also shown promising results in medical image computing. For example, Brosch and Tam [20] used Deep Belief Networks to learn features for 3-D brain image segmentation, Le et al. [84] used ISA for feature extraction from H&E histology images of glioblastoma multiforme, and Jurrus et al. [76] constructed a series of artificial neural networks to learn context information for large electron microscopy datasets. ISA as a feature learning method was also applied to prostate MR segmentation by Liao et al. [92]. However, in general learning feature extractors directly from data requires a relatively large number of training images and long training time.

4.3 Patch encoding

In the following, *patch* denotes a cube-shaped OPT image region to be classified; *window* denotes a smaller cube-shaped region of d^3 voxels from which local texture features are extracted. Figure 4.1 shows example patches from LGD, HGD and ICA. LGD tends to have somewhat regular texture with low spatial frequency as shown in Figure 4.1 (a); LGD normally has tubular morphological structure. ICA, shown in Figure 4.1 (c), contains more homogeneous micro-texture patterns corresponding to more dense tissue. HGD is intermediate in appearance as shown in Figure 4.1 (b). The bag-of-words framework is adopted to encode patches with a pre-learned visual dictionary [155]. The procedure of encoding a patch is illustrated in Figure 4.2. A window slides through the patch and at each location, a texture feature vector is extracted from the window. The feature vectors are quantised into visual words by

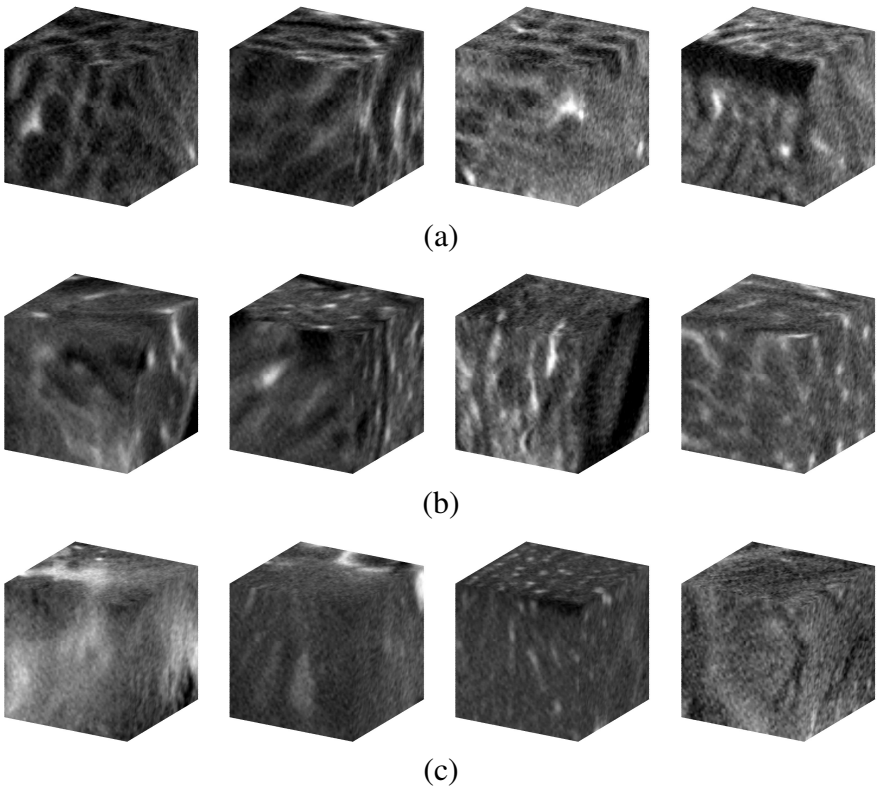


Figure 4.1 OPT image patches from regions labelled as (a) low-grade dysplasia (LGD), (b) high-grade dysplasia (HGD), and (c) invasive cancer (ICA).

matching with the most similar visual word in a pre-learned dictionary. Finally, an ℓ_1-normalised histogram of visual word frequencies constitutes the patch representation. Window step size was set to half the window width. Dictionaries of 200 visual words were obtained with k-means++ [12]. Bag-of-words provides a compact representation of fixed dimensionality regardless of the number of local windows used. It offers a uniform approach to a comparative evaluation of local feature extraction methods. The next section elaborates the feature extraction methods used.

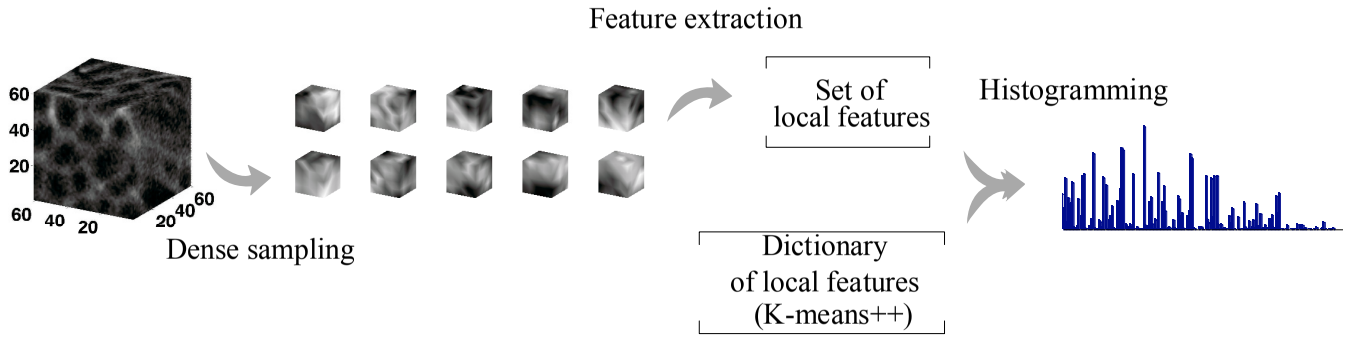


Figure 4.2 The procedure of encoding a patch as a bag-of-words histogram.

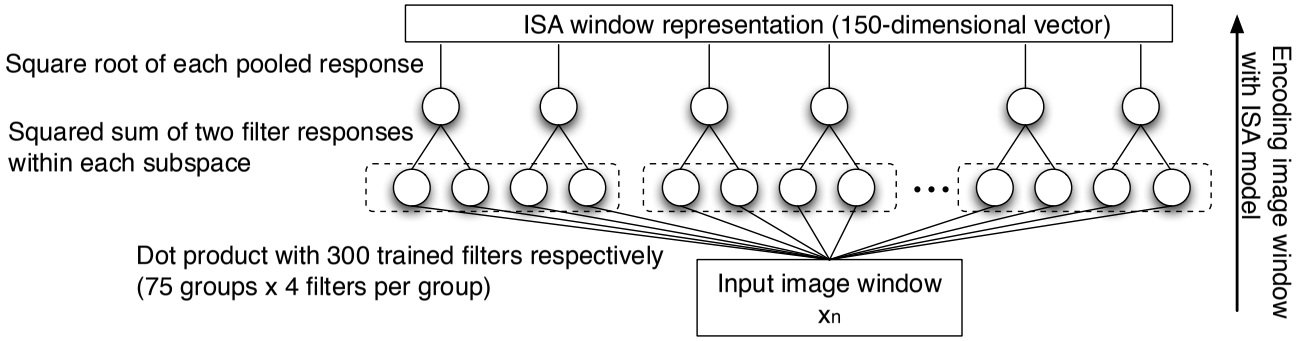


Figure 4.3 The procedure of encoding an image window with ISA model.

4.4 Random projection

Let $\mathbf{X}$ be a $d^3 \times N$ matrix in which the elements in each column are the voxel values of one of N windows. The RP method maps windows onto a k-dimensional subspace using a suitably generated $k \times d^3$ random projection matrix $\mathbf{R}$ (Equation (4.1)).

$$\hat{\mathbf{X}} = \mathbf{R}\mathbf{X}. \tag{4.1}$$

Each element in the matrix $\mathbf{R}$ is an independent sample from a standard normal distribution, i.e., a Gaussian with zero mean and unit variance. After projection, the columns in $\hat{\mathbf{X}}$ are considered as the window descriptors. According to the Johnson-Lindenstrauss lemma [33], data points in $\mathbb{R}^{d^3}$ are embedded into the lower-dimensional Euclidean space $\mathbb{R}^k$ such that pairwise distances between columns in $\mathbf{X}$ are approximately preserved. The computational complexity of RP is $O(d^3kN)$. In experiments reported in this thesis, k was set to 200 unless $d^3 < 200$ in which case the window was transformed with a square ($d^3 \times d^3$) random projection matrix.

4.5 3-D local binary patterns

Computing LBP from 2D images involves thresholding each 3×3-pixel neighbourhood at the value of its central pixel thus obtaining an 8-bit binary code. A histogram of these codes over an image window can then be used as a local descriptor. This representation, known as $\text{LBP}_{8,1}$, (i.e., 8-bit LBP with radius 1 neighbourhood), is not rotation invariant. Ojala et al. [111] found that the vast majority of binary codes in a local neighbourhood are so-called *uniform patterns* — the uniform appearance of the local binary pattern, i.e., there are a limited number of transitions or discontinuities in the circular presentation of the pattern. The most frequent “uniform" binary patterns correspond to primitive microfeatures, such as edges, corners, and spots. To achieve rotational invariance (around the central pixel) using uniform patterns, all non-uniform

LBP patterns are stored in a single bin in the histogram computation. The length of the uniform $LBP_{8,1}$ descriptor is 59 bins, which is smaller than the permutations (2^8 bins). In a 3-D volumetric image, designing LBPs that are invariant to arbitrary rotations is not straight-forward as the ordering of 3-D neighbourhood is undefined. Fehr and Burkhardt [45] addressed the problem by computing spherical correlations in the frequency domain. The approach is robust to 3-D rotations however it is computationally expensive. We choose to approximate the 3-D LBP with uniform $LBP_{8,1}$ descriptors computed in each of three orthogonal planes, taken to be aligned with the image axes for convenience [160]. This is computationally feasible in our experiments. In addition, this approximated 3-D LBP has been demonstrated to be competitive with a full volumetric LBP for encoding texture features in microscopic images [99]. The 3-D uniform $LBP_{8,1}$ operator encodes a window as a histogram with 177 bins.

4.6 Independent subspace analysis

ISA is an unsupervised feature learning method based on natural image statistics [73]. The features learned by ISA exhibit phase- and shift-invariant properties and have similarities to features computed by complex cells in the V1 area of primate visual cortex.

In the ISA method, image filters $\mathbf{W} = \{\mathbf{w}_t\}_{t=1}^{T}$ can be learned from a set of image windows $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$. Invariance is achieved by grouping the image filters into subspaces: the filter responses within the same subspace are pooled together, while among different groups the filter responses are treated independently. The model is

estimated by maximising the log-likelihood function [74]:

$$\log \mathcal{L}(\mathbf{X}|\mathbf{W}) = -\sum_{n=1}^{N}\sum_{l=1}^{L}\sqrt{\sum_{t\in g(l)}(\mathbf{w}_t^{\mathsf{T}}\mathbf{x}_n)^2}, \tag{4.2}$$

$$\text{subject to : } \mathbf{W}\mathbf{W}^{\mathsf{T}} = \mathbf{I}, \tag{4.3}$$

where each image filter $\mathbf{w}_t$ is a d^3-dimensional vector which is applied to the image window; $g(l)$ is the set of indices of lth group in the total L groups; $\mathbf{I}$ is an identity matrix. The constraint (4.3) is introduced to reduce the number of free parameters and leads to more stable solutions [72].

This optimisation was performed using stochastic gradient descent, after applying whitening transformations to remove correlations between voxels and then PCA to reduce the dimensionality to 300 (whenever the number of voxels in the window was greater than 300). A set of 300 filters forming 75 independent subspaces (4 filters per subspace) were *simultaneously* learned by optimising Formula (4.2). The parameters were chosen according to the empirical studies described in [90]. For each optimisation process, 200 passes of stochastic gradient descent with adaptive learning rate were applied through the entire set of training windows. After training, windows are represented as vectors containing 300 filter responses. Some filters learned from the training window set are visualised in Figure 4.4. These 300 responses are further pooled into a 150-dimensional vector as the final window representation. The pooling process takes the square root of the sum of squared two responses in the same subspace (illustrated in Figure 4.3). Filters within a given subspace are similar.

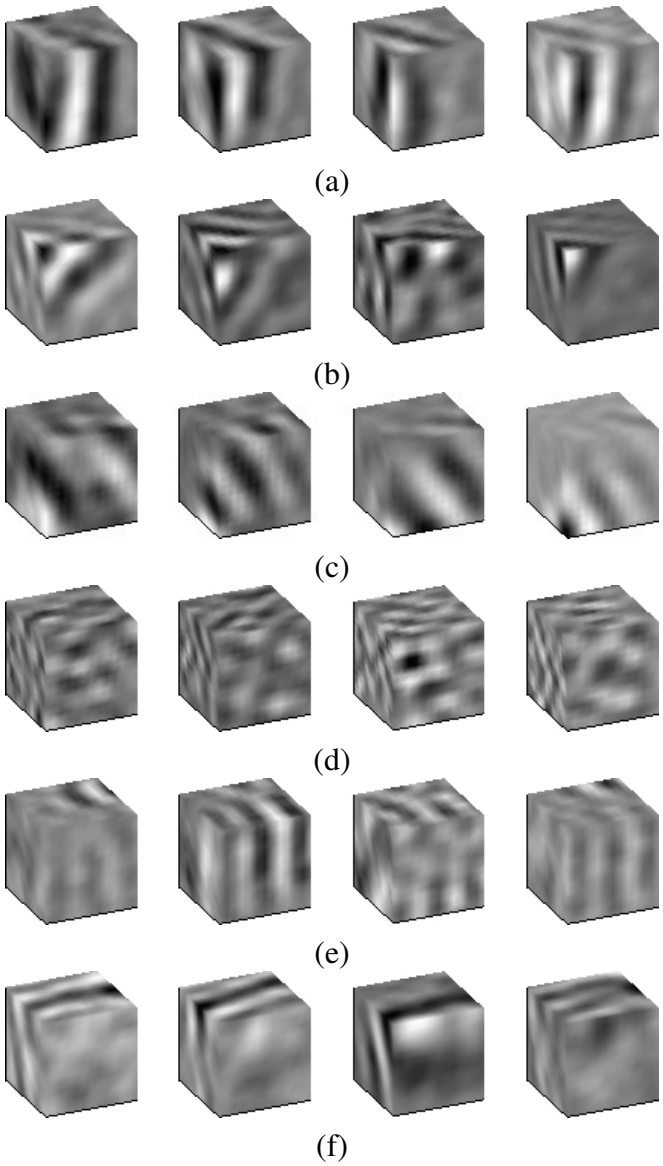


Figure 4.4 (a-f) Six groups of 4 filters learned from 2,700 windows (filter size: $13 \times 13 \times 13$). Filters within the same subspace (group) shared some similarities while filters in different subspaces had different orientations and frequencies.

4.7 Summary

This chapter described the rationale of the choices of texture analysis methods and provided technical details of the settings. In the next chapter, these methods will be used to encode OPT image patches, and compared and contrasted in classification and ordinal regression formulations.

Chapter 5

3-D patch classification and ordinal regression

5.1 About this chapter

The previous chapter described three representative texture feature extraction methods that have shown promising results in the literature. This chapter provides a rigorous evaluation of these methods for the task of discriminating between OPT patches of LGD, HGD and ICA. A classification model and an ordinal regression model both based on margin maximisation are applied in the experiments. The issues of class imbalance and output calibration are investigated empirically. Two strategies for fast approximation of non-linear kernels are also evaluated.

5.2 Multi-class classification

When formulated as a three-class classification problem, the task of discriminating OPT patches was addressed using a set of three SVM binary classifiers [5]. Each classifier was trained to discriminate one class from the others. This one-vs-rest approach leads to unbalanced datasets for each of the classifiers. Furthermore, it is important to calibrate

the outputs of the classifiers before comparing them in order to infer the class label. A further consideration is the use of non-linear kernels in the classifiers which, given large datasets, must be approximated for practical reasons. This chapter describes the classification method used and considers available solutions to each of these issues.

5.2.1 Binary subproblem

Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$ where $\mathbf{x}_i$ is the i^{th} bag-of-words feature vector and $y_i \in \{-1, 1\}$ is its corresponding class label, a classification function f that maps the feature vector to the label set $\{-1, 1\}$ can be found using the widely used SVM formulation by minimising

$$L(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i, \tag{5.1}$$

$$\text{subject to : } y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \tag{5.2}$$

where C is a parameter controlling the trade-off between model complexity and training errors; $\xi_i,\ i = 1, \ldots, l$ are slack variables; $\mathbf{w}$ is the weight vector, b is a bias weight, and $f(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x} + b$ is the SVM hyperplane to be optimised. This optimisation was performed using the primal form solver provided in the LIBLINEAR package [44]. C was searched over the set $\{2^\lambda \,|\, \lambda \in \mathbf{Z} \text{ and } \lambda \in [-15, 15]\}$. We observed that the classification accuracies were not very sensitive to the choice of C. $C = 0.1$ usually gave high accuracies.

5.2.2 Handling class imbalance

The binary classification subproblems do not have balanced datasets because in each case one class is being discriminated against all other classes. Class imbalance also arises because HGD is a less commonly assigned label than LGD or ICA in our dataset (see Figure 2.4). The classifier trained with the imbalanced data could overfit the

dominating class. This problem can be addressed by replacing the free parameter C with C_p and C_n for positive and negative classes respectively. Formula (5.1) then becomes:

$$\underset{\mathbf{w},\xi^+,\xi^-}{\text{Minimise}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_p \sum_{i=1}^{l^+} \xi_i^+ + C_n \sum_{i=1}^{l^-} \xi_i^-, \tag{5.3}$$

$$\text{subject to :} \quad \mathbf{w}^\intercal \mathbf{x}_i^+ + b \geq 1 - \xi_i^+, \tag{5.4}$$

$$-\mathbf{w}^\intercal \mathbf{x}_i^- - b \geq 1 - \xi_i^-, \tag{5.5}$$

$$\xi_i^+ \geq 0, \quad \xi_i^- \geq 0, \tag{5.6}$$

where $\mathbf{x}_i^+$, $\mathbf{x}_i^-$ are positive and negative training examples in the one-vs-rest setting; l^+ and l^- are the numbers of such examples; $\xi_i^+,\ i = 1, \ldots, l^+$ and $\xi_i^-,\ i = 1, \ldots, l^-$ are slack variables; $\mathbf{w}$ is the weight vector, b is a bias weight, and $f(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x} + b$ is the SVM hyperplane to be optimised; C_p and C_n are set as

$$C_p = \frac{l^+ + l^-}{2l^+} C_0, \tag{5.7}$$

$$C_n = \frac{l^+ + l^-}{2l^-} C_0. \tag{5.8}$$

C_0 was searched over the set $\{2^\lambda \mid \lambda \in \mathbf{Z} \text{ and } \lambda \in [-15, 15]\}$.

5.2.3 Output calibration

Three binary classifiers are trained for the three-class classification problem. Traditionally in the one-vs-rest scheme the final output score for a test example is the highest among the scores given by the three classifiers. However, since the three classifiers are trained independently, the scores are not necessarily comparable. This situation is helped by calibrating the scores prior to making this comparison by using Platt's scaling method to obtain values that can be treated as class probability estimates [116]. Calibration maps the binary classifier output ($\mathbf{w}^\intercal \mathbf{x}$) onto values that can be treated as

probabilities with a parameterised sigmoid function:

$$P(y = 1|\mathbf{x}; \mathbf{w}, A, B) = \frac{1}{1 + \exp(A\mathbf{w}^\intercal\mathbf{x} + B)}, \tag{5.9}$$

where A and B are learned from a validation set. For k-class classification, k such sigmoid functions are estimated, one per binary classifier. Unbalanced datasets mean that the fitting of these sigmoid functions is more heavily effected by over-represented classes. This problem was recently addressed by [146] who introduced bagging of under-sampling data estimators to refine the calibration procedure. Similarly, we apply the bagging method to one-vs-rest classifier calibration. More specifically, T balanced patch sets are formed by randomly discarding patches from over-represented classes and T calibration models are learned from these sets. The final probability estimate from a binary classifier is an average of the T models, i.e.,

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{T}\sum_{t=1}^{T} P^{(t)}(y = 1|\mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)}), \tag{5.10}$$

where $P^{(t)}(y = 1|\mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)})$ is obtained by applying Formula (5.9) with $A^{(t)}$ and $B^{(t)}$ estimated from the tth balanced patch set.

5.2.4 Non-linear kernel approximation

Using appropriate non-linear kernels that map feature vectors into a high-dimensional space can improve the classification performance of bag-of-words encoding. For example, the χ^2 kernel has been used for classification of endoscopic images [138]. However, this is computationally prohibitive for large-scale problems due to the expensive operation of constructing a Gram matrix over all training data. Recently the method of approximating kernels with explicit feature maps enabled the use of non-linear kernels with relatively low computational cost on large scale datasets [98, 145]. The main idea is that for a homogeneous kernel $k(\mathbf{x}, \mathbf{y}) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, it is possible to use an

approximation function $\Phi(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^w$ so that $k(\mathbf{x}, \mathbf{y}) \approx \Phi(\mathbf{x})^\intercal \Phi(\mathbf{y})$. The $\Phi(\cdot)$ can be data-independent. A linear SVM can then be used directly with $\Phi(x)$ for classification. This leverages the high performance of non-linear kernels while maintaining scalability to large scale problems.

Given two OPT patches encoded with bag-of-words, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the χ^2 kernel is computed as

$$k_{\text{chi}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} \frac{2x_i y_i}{x_i + y_i}, \tag{5.11}$$

and the histogram intersection kernel is computed as

$$k_{\text{hist}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} \min(x_i, y_i), \tag{5.12}$$

where x_i and y_i are the ith component of $\mathbf{x}$ and $\mathbf{y}$ respectively. These two kernels were approximated with methods proposed in [145] and [98] respectively.

5.3 Ordinal regression

The labels of dysplastic change are qualitative measurements that reflect increasing severity in diagnosis from LGD to HGD to ICA. This suggests that a patch label $r \in \{\text{LGD}, \text{HGD}, \text{ICA}\}$ could be an ordinal variable rather than a nominal one; the order is LGD $\prec$ HGD $\prec$ ICA.

5.3.1 Large-margin formulation

Given a training set $\{\mathbf{x}_i, r_i\}_{i=1}^{l}$ where $\mathbf{x}_i$ is the i^{th} bag-of-words feature vector and $r_i \in \{\text{LGD}, \text{HGD}, \text{ICA}\}$ is its corresponding ordinal label, a function g that maps the feature vector to the label set $\{\text{LGD}, \text{HGD}, \text{ICA}\}$ can be found using a structural risk minimisation formulation of ordinal regression. Herbrich et al. [69] show how to estimate g by learning a ranking function $f(\mathbf{x}) \in \mathbb{R}$ so that the pairwise orders are

preserved, i.e.,

$$f(\mathbf{x}_i) \prec f(\mathbf{x}_j) \iff r_i \prec r_j. \tag{5.13}$$

The score $f(\mathbf{x}_i)$ is thresholded to determine the ordinal category of $\mathbf{x}_i$, i.e.,

$$g(\mathbf{x}) = k \iff f(\mathbf{x}) \in [\theta_k, \theta_{k+1}] \tag{5.14}$$

where θ_k and θ_{k+1} are the learned thresholds of the kth category. In the case of a linear mapping $f(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$, the weight vector $\mathbf{w}$ can be learned by minimising the empirical risk on the pairwise order set $\mathcal{P} = \{(i, j)|r_i \prec r_j\}$:

$$L(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(i,j)\in\mathcal{P}} \xi_{ij}, \tag{5.15}$$

$$\text{subject to : } t_{ij}(\mathbf{w}^\intercal \mathbf{x}_j - \mathbf{w}^\intercal \mathbf{x}_i) \geq 1 - \xi_{ij},\ \xi_{ij} \geq 0,\ \forall (i, j) \in \mathcal{P},$$

where t_{ij} is the order label, i.e., $t_{ij} = 1$ iff $r_i \prec r_j$; $t_{ij} = -1$ otherwise; $\xi_{ij}, \forall (i, j) \in \mathcal{P}$ are slack variables.

There are three types of pairwise ordering in $\mathcal{P}$, i.e., LGD $\prec$ HGD, LGD $\prec$ ICA and HGD $\prec$ ICA (the number of pairs of each type is denoted as N_1, N_2, and N_3 respectively). Similar to the balancing of multi-class classification in Section 5.2.2, the balance problem is also considered on the pairwise order set. Specifically, we replace C with three cost parameters to be proportional to the number of pairwise preferences of each type respectively, i.e.,

$$C_{\text{LGD}\prec\text{HGD}} = \frac{C_0(N_1 + N_2 + N_3)}{N_1}, \tag{5.16}$$

$$C_{\text{LGD}\prec\text{ICA}} = \frac{C_0(N_1 + N_2 + N_3)}{N_2}, \tag{5.17}$$

$$C_{\text{HGD}\prec\text{ICA}} = \frac{C_0(N_1 + N_2 + N_3)}{N_3}, \tag{5.18}$$

where C_0 is a free parameter; the value is searched within the set $\{2^i \mid i \in \mathbf{Z} \text{ and } i \in [-15, 15]\}$.

5.3.2 Solving the optimisation problem

The objective function in Formula (5.15) is a standard quadratic programming problem that can be solved by many existing convex optimisation packages. However, the size of the pairwise order set grows quadratically with the number of samples, e.g., 100 LGD and 100 ICA patches gives $10,000$ pairwise orders. Most solvers are not feasible due to the scale of our problem. Here we choose the fast rank SVM solver proposed in [24] which tackles the primal form of rank SVM with Newton's method. Instead of computing the inverse Hessian matrix in the Newton step, the fast rank SVM approximates the inverse with a conjugate gradient method. This approximation is both fast and memory efficient.

With the estimated $\hat{\mathbf{w}}$, the optimal threshold θ_k is set to be in the middle of the closest correctly separated training pair in the kth and $(k+1)$th category, i.e.,

$$\theta_k = \frac{\hat{\mathbf{w}}^\top \mathbf{x}_{i_k} + \hat{\mathbf{w}}^\top \mathbf{x}_{j_k}}{2}, \tag{5.19}$$

$$\text{where:} \quad (i_k, j_k) = \underset{(i,j)\in\mathcal{P}_k}{\operatorname{argmin}}(\hat{\mathbf{w}}^\top \mathbf{x}_j - \hat{\mathbf{w}}^\top \mathbf{x}_i), \tag{5.20}$$

$$\mathcal{P}_k = \{(i, j) | r_i = k \wedge r_j = k + 1 \wedge (\hat{\mathbf{w}}^\top \mathbf{x}_j - \hat{\mathbf{w}}^\top \mathbf{x}_i) \geq 1\}. \tag{5.21}$$

5.4 Patch sampling and cross-validation

Empirical evaluations used patches sampled with a systematic uniform random sampling (SURS) strategy. The major advantage of SURS over repeated uniform random sampling is its statistical efficiency. The sampling method is illustrated in Figure 5.1. The displacement between points on the sampling grid was set to 31 voxels along each of the three image axes. Patches were sampled with their centres at the grid points.

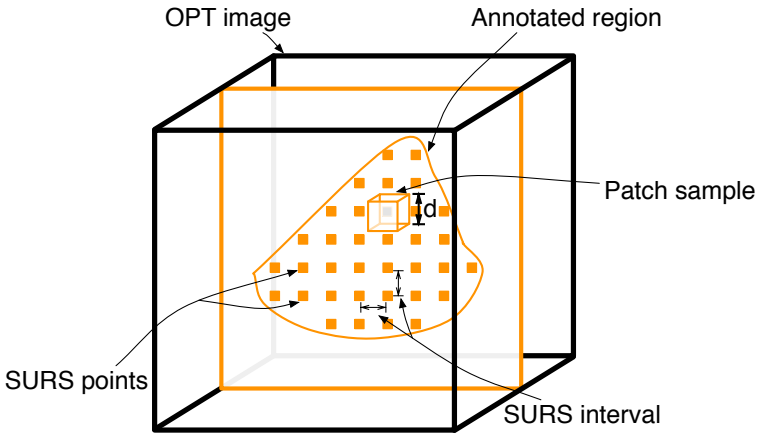


Figure 5.1 Patch sampling with SURS applied to an annotated region. A regular sampling grid with fixed displacement between sampling points is positioned at random. Patches are sampled with their centres at the grid points.

A set of 20,000 patches was sampled from the 90 polyps at each of 10 patch sizes, $d \in \{11, 21, 31, \ldots, 111\}$, giving 200,000 patches in total. This enabled exploration of the effect of patch size as a system parameter.

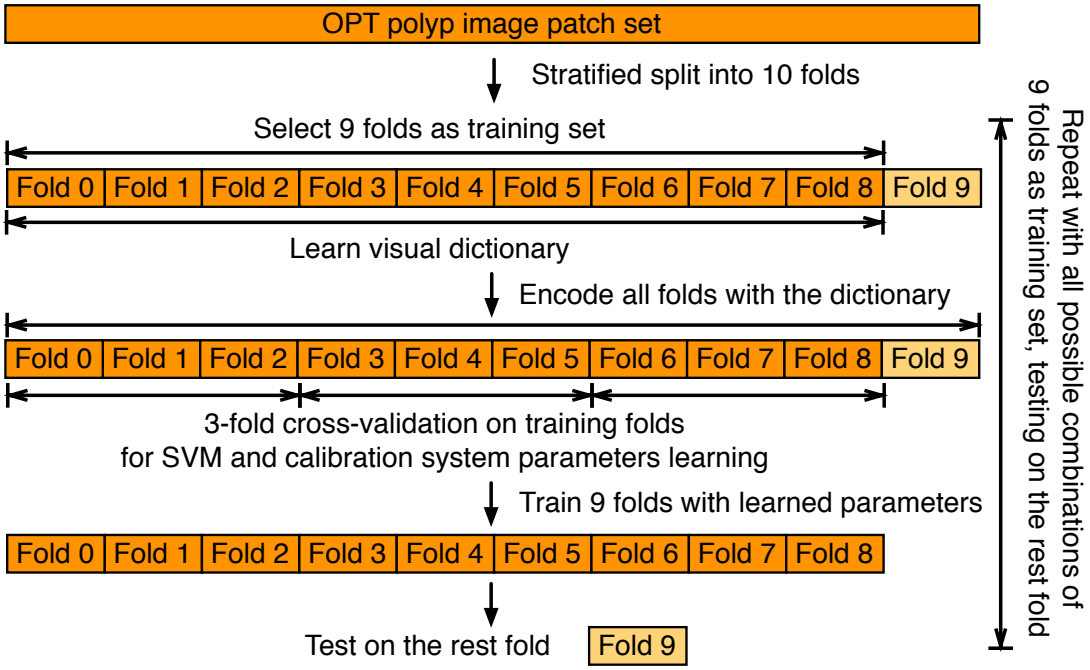


Figure 5.2 Cross-validation scheme.

Experiments based on 10-fold cross-validation with stratified splitting were carried out to estimate generalisation capability of the systems as follows. (1) All patches from the same image are in the same fold; this ensures that no part of the polyp participating

in the test is in the training set (cross-polyp generalisation); (2) the testing fold is generated so that it always contains the same proportions of polyp images in the three classes as the original dataset to ensure that there is no dominating class in the test dataset. Within the training folds, we further applied 3-fold cross-validation to search the appropriate system parameters (e.g., SVM calibration parameters, cost parameters). This experimental design is illustrated in Figure 5.2. The results in the following section are based on this cross-validation scheme unless otherwise specified.

5.5 Performance metrics

There are six types of mis-classification, i.e., LGD as ICA, LGD as HGD, ICA as HGD, ICA as LGD, HGD as LGD, HGD as ICA. Error rates were calculated for each error type respectively. Performance was also compared with measures of overall mis-classification rate, absolute error rate and F-measure. Overall mis-classification rate is the number of mis-classified cases divided by the total number of test cases, N, without considering ordinal information. The absolute error rate is $(\sum_{n=1}^{N} |e_n|)/N$, where e_n is a scalar error value. In the case of correct prediction, $e_n = 0$. In the case of an out-by-one error (LGD confused with HGD or HGD confused with ICA), $e_n = 1$. In the case of an out-by-two error (LGD confused with ICA) $e_n = 2$. For experiments on handling class imbalance, averaged F-measure was used. The averaged F-measure is an average over F-measures with respect to each class. The 95% confidence intervals of each type of measurement were obtained by bootstrapping [38] with $n = 10,000$; specifically we generated n bootstrap replicates of the classifier outputs, and calculated average F-measures of each bootstrap replicate. The confidence intervals were computed with n average F-measures using the `boot.ci` function from the `boot` package (`R` implementation). When evaluating window and patch size parameters, and kernel approximations, performance was measured with Averaged Area under the ROC curve (AAUC) which is not affected by specific choices of thresholds on the raw SVM

outputs. Error bars indicating 95% confidence intervals were estimated with the method proposed by [35]. When comparing features with ordinal regression formulations, ROC surfaces were constructed with true positive rates. Volumes under the surfaces were further calculated using the algorithm proposed by [97].

5.6 Results

5.6.1 Overall comparison of formulations and feature types

Table 5.1 Multi-class classification (left) and ordinal regression (right) confusion matrices.

((a)) Random projection features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.713	0.212	0.075
	HGD	0.298	0.528	0.174
	LGD	0.088	0.130	0.782

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.535	0.418	0.048
	HGD	0.228	0.556	0.216
	LGD	0.025	0.275	0.701

((b)) Local binary pattern features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.722	0.168	0.110
	HGD	0.340	0.402	0.258
	LGD	0.081	0.119	0.800

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.417	0.491	0.093
	HGD	0.167	0.568	0.265
	LGD	0.031	0.244	0.725

((c)) Independent subspace analysis features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.628	0.217	0.155
	HGD	0.273	0.481	0.247
	LGD	0.084	0.192	0.724

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.541	0.362	0.097
	HGD	0.236	0.530	0.234
	LGD	0.043	0.339	0.618

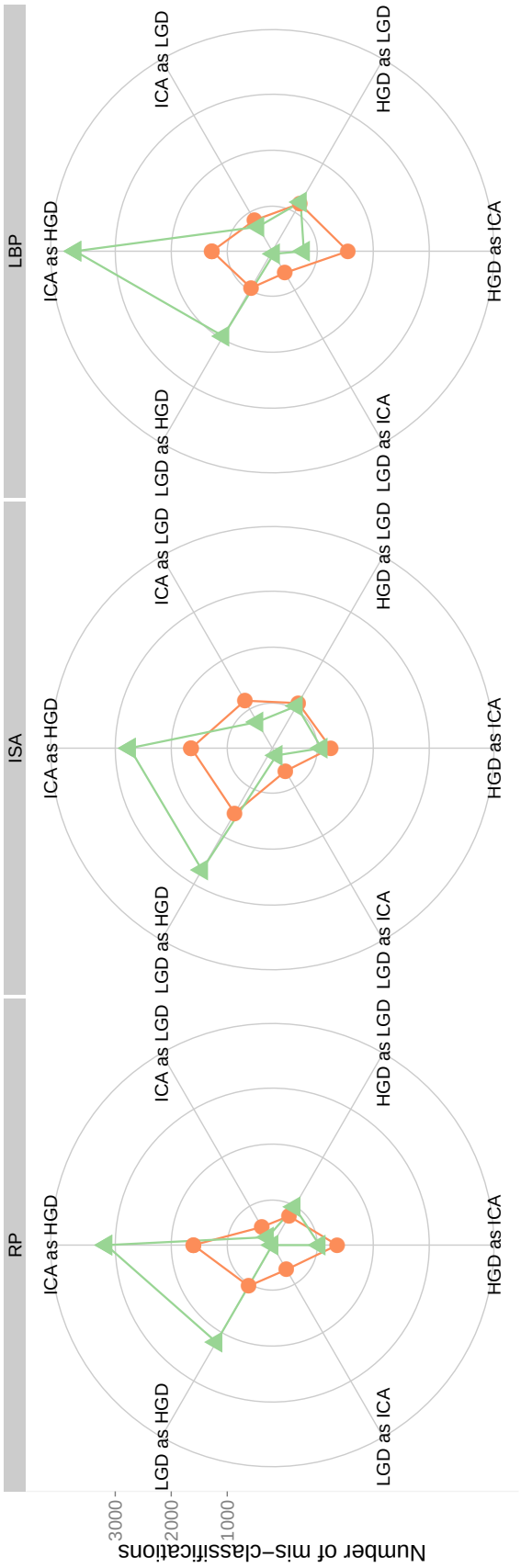


Figure 5.3 Cobweb diagrams showing number of mis-classifications for multi-class classification and ordinal regression formulations.

Table 5.2 Multi-class classification (Mult.) and ordinal regression (Ordi.) results for different features (with 95% confidence intervals).

Type	Mis-classification rate	Absolute error rate	Average F-measure
RP - Mult.	**0.304** (0.297, 0.310)	**0.366** (0.358, 0.374)	**0.668** (0.661, 0.674)
ISA - Mult.	0.352 (0.346, 0.359)	0.431 (0.422, 0.440)	0.617 (0.611, 0.624)
LBP - Mult.	0.346 (0.339, 0.352)	0.423 (0.414, 0.431)	0.622 (0.615, 0.628)
RP - Ordi.	**0.425** (0.419, 0.432)	**0.448** (0.441, 0.456)	**0.574** (0.567, 0.580)
ISA - Ordi.	0.440 (0.433, 0.447)	0.497 (0.488, 0.505)	0.554 (0.548, 0.561)
LBP - Ordi.	0.428 (0.421, 0.435)	0.473 (0.465, 0.481)	0.560 (0.553, 0.566)

Table 5.3 Multi-class classification results for different classifiers and features (with 95% confidence intervals).

Type	Mis-classification rate	Absolute error rate	Average F-measure
RP - RF.	**0.311** (0.305, 0.317)	**0.402** (0.394, 0.411)	**0.624** (0.617, 0.631)
ISA - RF.	0.337 (0.331, 0.344)	0.418 (0.409, 0.427)	0.599 (0.593, 0.606)
LBP - RF.	0.323 (0.317, 0.330)	0.420 (0.411, 0.429)	0.591 (0.585, 0.598)
RP - KNN.	**0.357** (0.351, 0.364)	0.467 (0.458, 0.476)	**0.603** (0.596, 0.609)
ISA - KNN.	0.362 (0.355, 0.369)	**0.467** (0.457, 0.476)	0.583 (0.576, 0.590)
LBP - KNN.	0.365 (0.358, 0.372)	0.477 (0.467, 0.486)	0.577 (0.570, 0.584)
RP - MLP.	**0.303** (0.297, 0.310)	**0.372** (0.363, 0.380)	**0.656** (0.649, 0.663)
ISA - MLP.	0.337 (0.331, 0.344)	0.414 (0.405, 0.422)	0.612 (0.605, 0.618)
LBP - MLP.	0.346 (0.340, 0.353)	0.462 (0.452, 0.471)	0.598 (0.591, 0.605)
RP - KSVM.	**0.299** (0.292, 0.305)	**0.364** (0.356, 0.374)	**0.675** (0.668, 0.681)
ISA - KSVM.	0.323 (0.316, 0.329)	0.385 (0.377, 0.393)	0.648 (0.642, 0.655)
LBP - KSVM.	0.300 (0.293, 0.306)	0.366 (0.357, 0.374)	0.664 (0.657, 0.670)

Figure 5.3 shows cobweb diagrams for multi-class classification and ordinal regression for each of the three feature types. Table 5.1 reports confusion matrices. Table 5.2 reports the misclassification rates, absolute error rates, and average F-measures. These results were obtained using the patch size $81 \times 81 \times 81$ and the window size $13 \times 13 \times 13$ using the procedures described in Sections 5.2.2 and 5.2.3. Experiments exploring the effect of varying such parameters are reported in the following section. The values in Table 5.2 suggest that RP outperforms ISA and LBP; the performance rank of RP is consistent for all measures considered for both problem formulations. Compared to multi-class classification, ordinal regression makes less confusion between the ICA and LGD classes. The ordinal regression formulation is a better choice over one-vs-rest classification when the focus concentrates on minimising the risk of mis-classification between LGD and ICA.

Table 5.3 compares classification performance of the one-vs-rest SVM classifications with three inherently multi-class classifiers: random forest (RF) [18], k-nearest neighbours (KNN), and multi-layer perceptrons (MLP). A non-linear classification function $\hat{f}$ that maps the feature vector to the label set $\{\text{LGD}, \text{HGD}, \text{ICA}\}$ was trained for each method to discriminate three classes. For RF classifier we report performances of 2,000 randomised classification trees. k was searched over the set $\{5, 10, 20\}$ in the KNN classifier. We set one hidden layer with 20 neurons for the MLP[1]. For the one-vs-rest SVM in the comparison, the procedures described in Section 5.2.2, Section 5.2.3 and Section 5.2.4 were used (denoted as KSVM). The non-linear kernel in KSVM was computed as the 1400-dimensional χ^2 approximation (the *Chi_2* method in Figure 5.9) described in Section 5.2.4. In most cases, RF, KNN and MLP performed worse than one-vs-rest KSVM. MLP gave the most competitive performance among the three inherently multi-class classifiers.

[1]We used the `newpr` function from the Matlab Neural Network Toolbox. (*URL:* http://uk.mathworks.com/products/neural-network/)

5.6.2 One-vs-rest classification

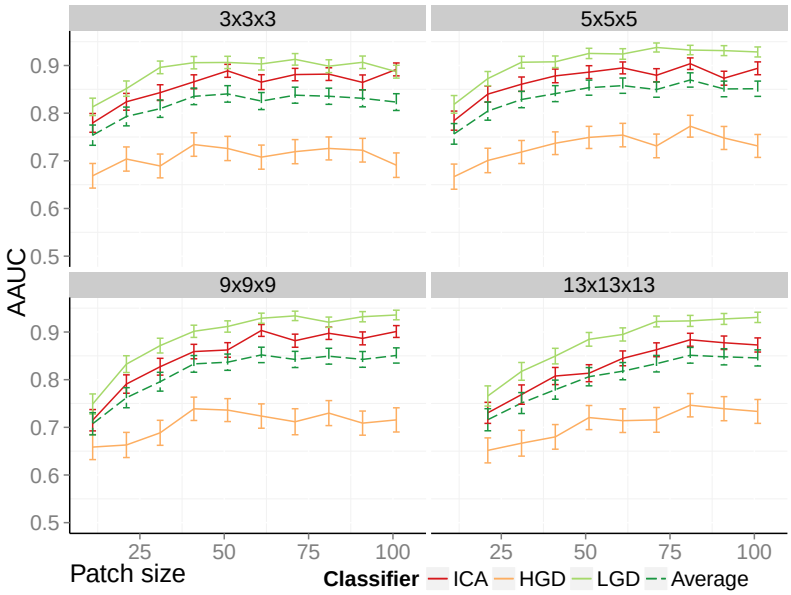


Figure 5.4 AAUC using random projection features with varied window size and patch size. Error bars show 95% confidence intervals.

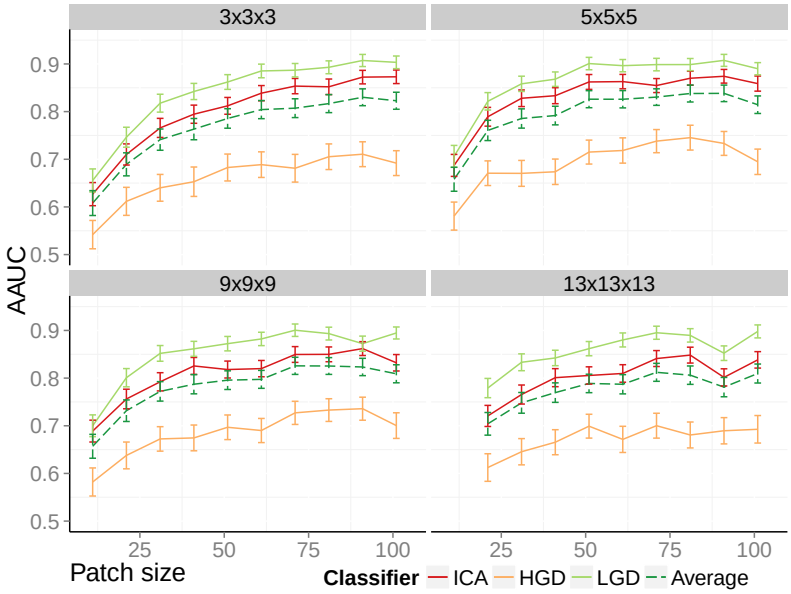


Figure 5.5 AAUC using local binary pattern features with varied window size and patch size. Error bars show 95% confidence intervals.

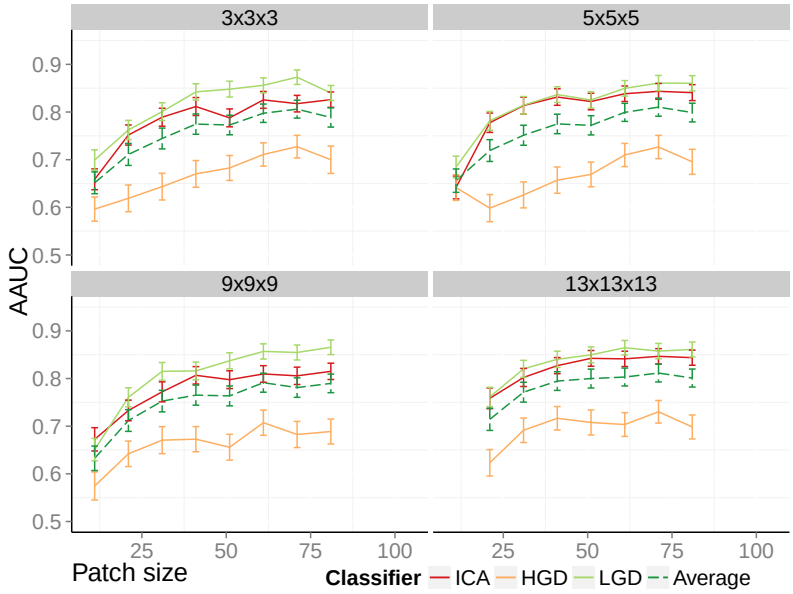


Figure 5.6 AAUC using independent subspace analysis features with varied window size and patch size. Error bars show 95% confidence intervals.

Figures 5.4, 5.5, and 5.6 plot averaged area under the ROC curves (AAUCs) averaged over 10 folds for each one-vs-rest classifier for various window and patch sizes. These plots suggest that patch size has greater impact than window size. At patch size $81 \times 81 \times 81$ all features reached high performance. At large window size ($13 \times 13 \times 13$) we generated a smaller number of windows than with smaller window sizes; this was computationally more efficient without significantly decreasing the AAUC. These parameter values were used in other experiments unless otherwise specified. RP gave generally good results compared to the other two types of feature.

Classifier calibration

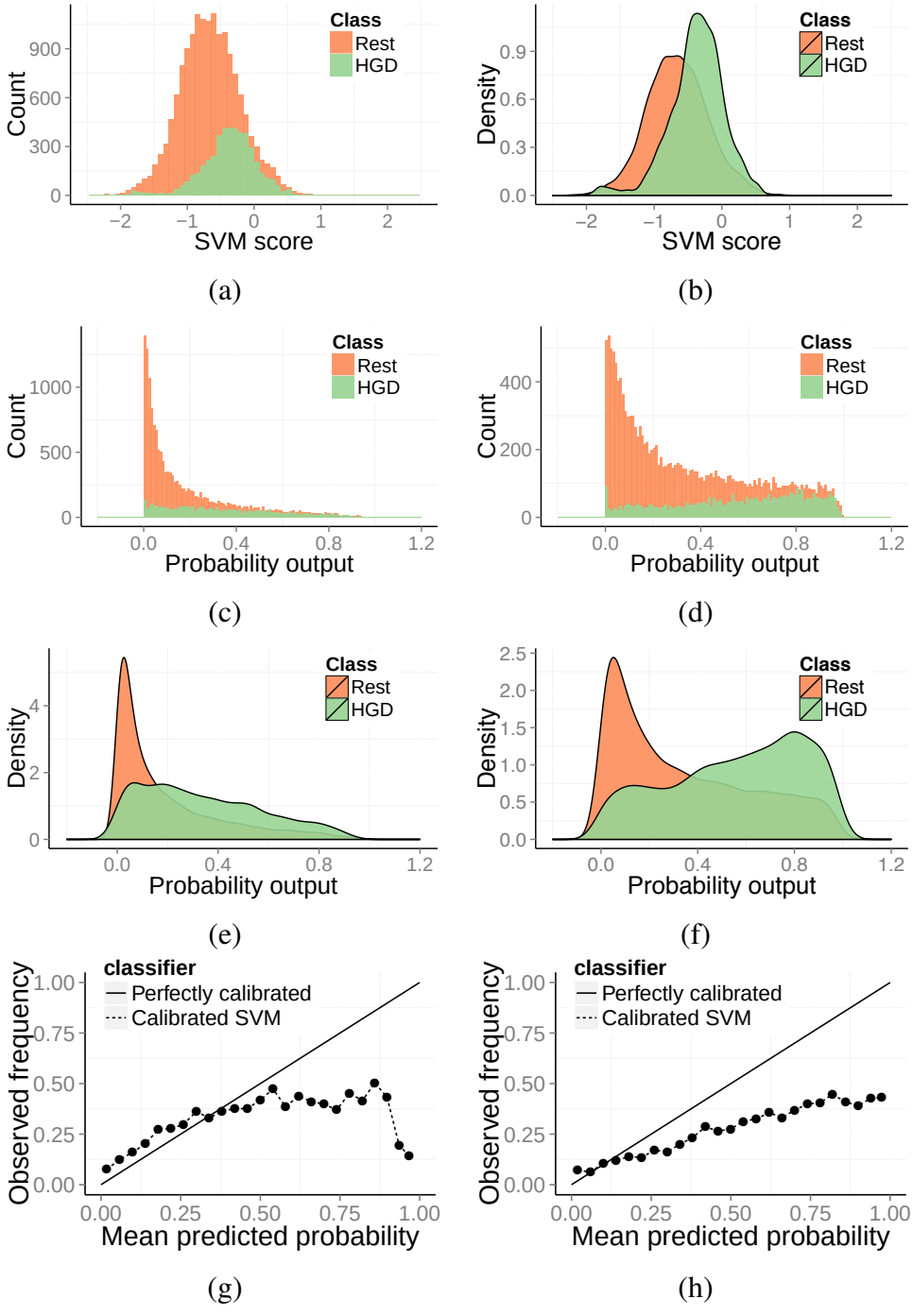


Figure 5.7 Calibrating the HGD-vs-rest classifier output of all testing folds. Figure best viewed in colour. (a) Histogram and (b) Gaussian kernel density estimates of raw SVM scores. (c,e,g) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (5.9). (d,f,h) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (5.10). The diagonal solid lines in the reliability diagrams indicate perfectly calibrated classifiers.

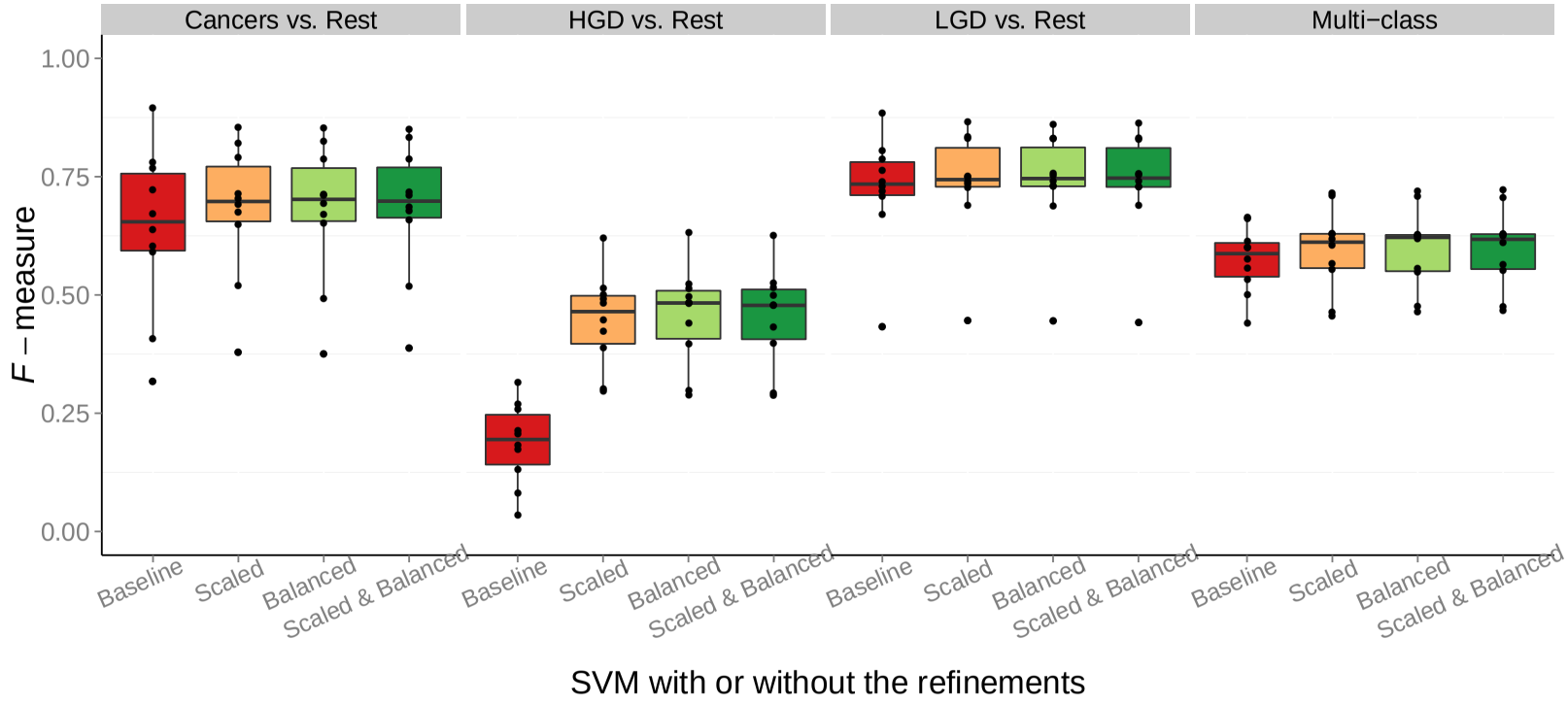


Figure 5.8 Box plots summarising the distribution of F-measures over cross-validation folds with ISA features (patch size: $81 \times 81 \times 81$, window size: $13 \times 13 \times 13$). *Balanced* denotes use of the balanced cost function. *Scaled* denotes use of output calibration. *Scaled and Balanced* denotes use of both a balanced cost function and output calibration.

Figure 5.7 shows calibration of an HGD-vs-rest SVM classifier. Thresholding the raw output scores at 0 gave an F-measure of 0.213 with 95% confidence interval (0.199, 0.226). After calibration using formula (5.9) (see Figure 5.7(c-e)), thresholding probabilities at 0.5 gave an improved F-measure of 0.329 with 95% confidence interval (0.309, 0.337). However, the probability of HGD is systematically underestimated by the sigmoid function in Formula (5.9) because the model is biased towards the "rest" class due to the imbalanced training set. This can be seen from Figures 5.7 (a-e). In Figures 5.7 (f–h) we visualise the probability outputs of the bagging method. The under-sampling of the dominating class and the ensemble strategy mitigated the bias problem. In the reliability diagrams (Figures 5.7 (e, h)), the observed probabilities are grouped into 25 bins and the observed frequency of positives are plotted against mean probability in each bin. For a perfect reliability, the observed frequency and the predicted probability should be equal (shown as dotted line in diagonal direction). By using the bagging method, the probability outputs from 0.8 to 1.0 are more reliable compared to the original Platt's method. We are able to further improve the F-measure from 0.329 to 0.464 with 95% confidence interval (0.453, 0.475). In the other experiments reported here we applied Equation (5.10) for the calibration procedure.

Effect of class balancing and calibration

We conducted a group of experiments to investigate the effect of balancing the cost function (Section 5.2.2) and output calibration (Section 5.2.3) on classification performance. As a *baseline* we use the standard SVM (Formula (5.1)). To obtain decisions from one-vs-rest classifiers, raw SVM scores were thresholded at 0 and calibrated outputs at 0.5. Multi-class classification decisions were obtained by applying all the one-vs-rest classifiers and predicting the class for which the corresponding classifier reports the highest score (max-win strategy). Figure 5.8 shows comparisons with the baseline SVM in terms of F-measures of each fold. Classification results after class balancing and calibration were significantly different from the baseline ($p < 0.0001$;

McNemar's test [39]). Use of the balanced SVM cost function generally improved performance, especially for HGD-vs-rest classification where the baseline suffers from the very unbalanced dataset (see Figure 2.4). SVM output calibration helped address the data imbalance problem and improved performance compared to the baseline. Combining the balanced cost function and output calibration does not lead to further improvement in terms of F-measure as compared to each applied individually.

Effect of kernel approximation

The effect of approximating the histogram intersection and χ^2 kernels as discussed in Section 5.2.4 was evaluated. Two approximation strategies were compared: (1) B-spline approximations[2] [98] of the χ^2 kernel (denoted as *Chi_1*) and the histogram intersection kernel (denoted as *Min_1*); and (2) the homogeneous feature map[3] [145] for the χ^2 kernel (denoted at *Chi_2*) and the histogram intersection kernel (denoted as *Min_2*). Figure 5.9 gives AUC results using those approximations compared with the baseline result of the plain linear SVM (error bars indicate 95% confidence intervals). It can be seen that, generally, the approximations of non-linear kernels can improve dysplasia classification over the baseline for all three feature types if the feature dimensionality is carefully tuned; the exception was that the performance improvement of B-spline approximations of the histogram intersection kernel was not stable and conclusive. The homogeneous feature map consistently outperformed the B-spline approximation although the improvement was marginal. The χ^2 kernel worked better than the histogram intersection kernel in most cases.

[2]We used the Matlab implementation by [98]. (*URL:* http://ttic.uchicago.edu/~smaji/projects/libspline-release1.0.tar.gz)

[3]We used the Matlab implementation provided in VLFeat package [144].

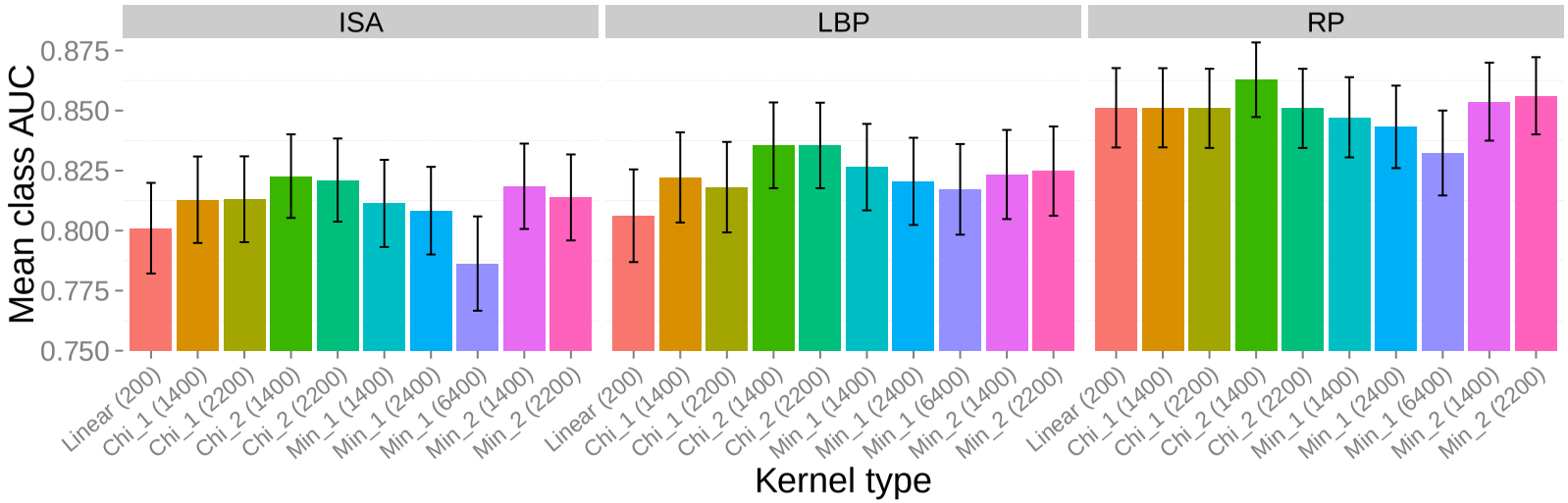


Figure 5.9 Mean class AUC for different kernel types (with feature dimensionality given in parentheses). Features were extracted with patch size $81 \times 81 \times 81$ and window size $13 \times 13 \times 13$. "Chi_1" and "Min_1" denote B-spline based kernel approximation of χ^2 and histogram intersection kernels respectively. "Chi_2" and "Min_2" denote homogeneous feature map of χ^2 and histogram intersection kernels respectively. Error bars indicate 95% confidence intervals.

Polyp classification

Table 5.4 Polyp classification results for different features (with 95% confidence intervals).

Type	Mis-classification rate	Absolute error rate	Average F-measure
RP	**0.411** (0.285, 0.538)	**0.344** (0.247, 0.443)	**0.655** (0.561, 0.756)
ISA	0.489 (0.361, 0.619)	0.422 (0.318, 0.525)	0.579 (0.482, 0.683)
LBP	0.478 (0.350, 0.610)	0.400 (0.300, 0.501)	0.595 (0.497, 0.700)

Table 5.5 Polyp classification confusion matrices.

((a)) Random projection features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.733	0.167	0.100
	HGD	0.167	0.567	0.267
	LGD	0.100	0.233	0.667

((b)) Local binary pattern features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.667	0.200	0.133
	HGD	0.300	0.433	0.267
	LGD	0.100	0.200	0.700

((c)) Independent subspace analysis features.

		Predictions		
		ICA	HGD	LGD
Labels	ICA	0.500	0.433	0.067
	HGD	0.233	0.533	0.233
	LGD	0.133	0.167	0.700

Table 5.4 reports multi-class classification results at polyp-level. These results were obtained using the window size $13 \times 13 \times 13$, and the procedures described in Sections 5.2.2 and 5.2.3. Instead of patch-level classification, each *polyp* was encoded using three types of local features and classified based on 10-fold cross-validation described in Section 5.4. Table 5.5 shows the confusion matrices. The values in Table 5.4 and 5.5 suggest that RP outperforms ISA and LBP in terms of all the measures we adopted.

5.6.3 Ordinal regression

Figure 5.10 shows distributions of rank SVM scores obtained along with the two thresholds for three-class ordinal regression using Formula (5.19). By varying the two

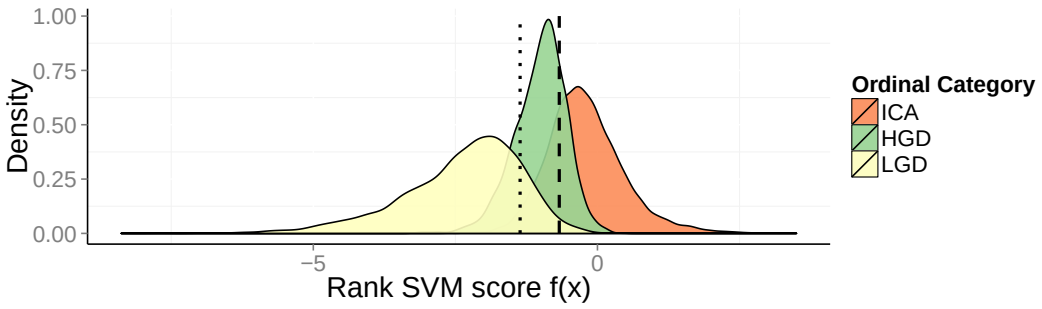


Figure 5.10 Distributions of training set ranking scores. Dashed and dotted vertical bars are thresholds for the three ordinal categories, obtained using Formula (5.19).

thresholds ROC surfaces can be created. Each point on an ROC surface represents a tuple of three true positive rates (TPR). Figure 5.11(a) shows ROC surfaces for the three feature types.

For visual comparison, each surface was mapped onto the plane $\mathrm{TPR}_{\mathrm{LGD}} + \mathrm{TPR}_{\mathrm{ICA}} + \mathrm{TPR}_{\mathrm{HGD}} = 1$ and the signed distances between points on the surface and the plane were colour-coded (Figure 5.11(b)). The plane represents the ROC surface for random classification. The signed distances give an indication of how much better the classifier is than random guessing. The signed distance maps for different feature types were further compared (Figure 5.11(c)). Compared to ISA features, RP and LBP give better performances in high TPR range of LGD (high LGD specificity range), whereas ISA is better in high TPR range of ICA (high ICA specificity range). In the comparison of RP and LBP, RP performs better when true positive rate of HGD is in low range (in low HGD specificity).

The volume under an ROC surface (VUS) is the expected proportion of correctly ranked triplets uniformly drawn from all possible samples of triplets [68]. It is an extension of area under the ROC curve to the three-class case. The VUS and its 95% confidence interval for the RP, ISA and LBP methods were 0.590 (0.582, 0.598), 0.522 (0.514, 0.529) and 0.533 (0.525, 0.541) respectively. In this experiment RP features showed significantly better performance than ISA and LBP in terms of VUS.

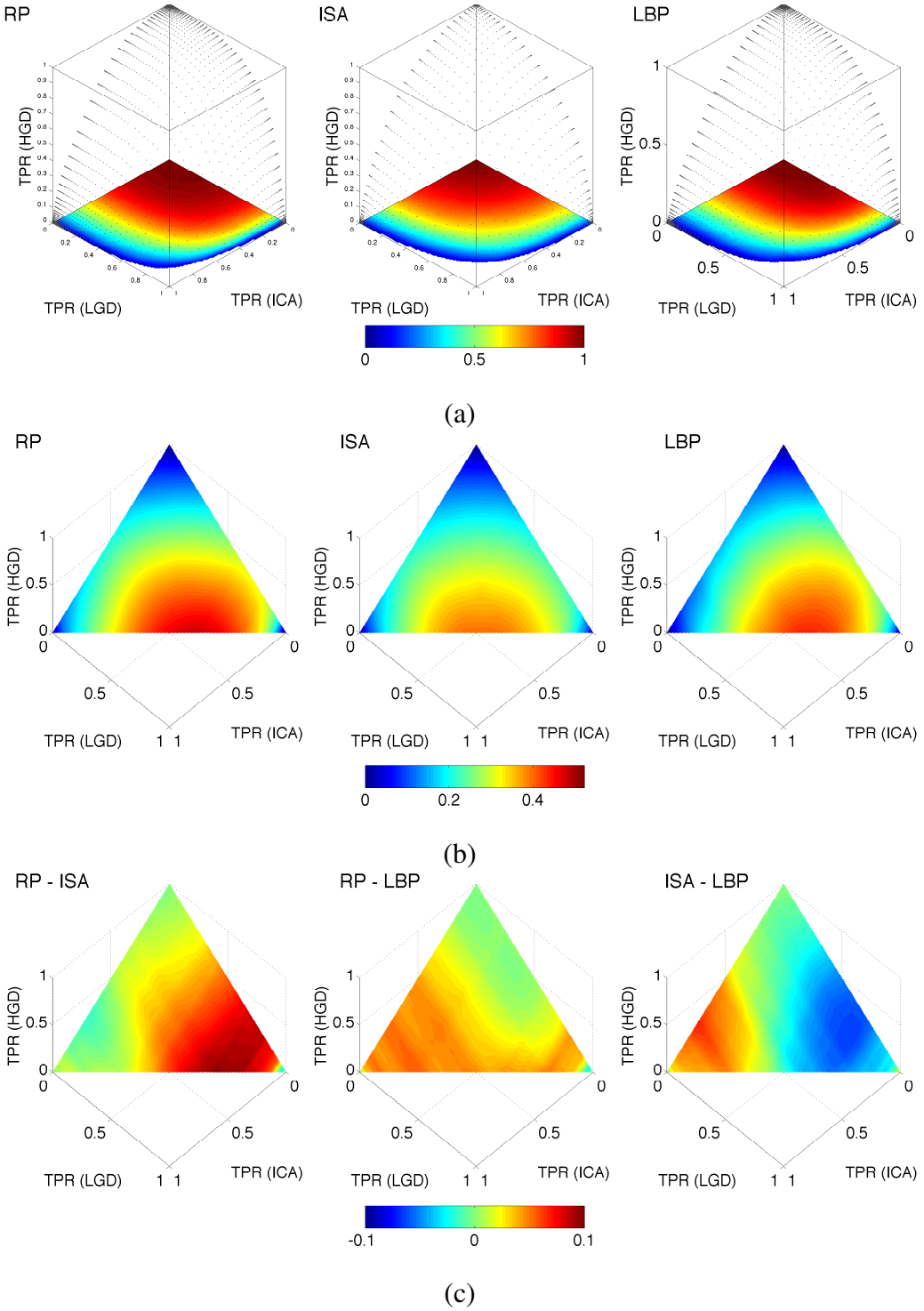


Figure 5.11 (a) ROC surfaces for three-class ordinal regression (patch size: $81 \times 81 \times 81$, window size: $13 \times 13 \times 13$). Colour on the $\mathrm{TPR}_{\mathrm{LGD}} - \mathrm{TPR}_{\mathrm{ICA}}$ plane encodes $\mathrm{TPR}_{\mathrm{HGD}}$ value. (b) Signed distances from points on the ROC surfaces to the plane $\mathrm{TPR}_{\mathrm{LGD}} + \mathrm{TPR}_{\mathrm{ICA}} + \mathrm{TPR}_{\mathrm{HGD}} = 1$. (c) Differences between the maps in (b).

5.7 Summary

Large margin multi-class classification and ordinal regression techniques were adopted with a focus on practical considerations. Their performance with three types of texture descriptor was demonstrated. The results showed that random projection features and bag-of-words framework was the best among the three methods. Although the RP method is relatively simple, it outperformed carefully designed LBP operators as well as automatically learned ISA filters for all parameter settings. Similar observations were also reported in several general texture classification systems [94, 143] where classification of filter-based features did not outperform raw pixels as patch representation. The RP method generates a low-dimensional representation without making strong assumptions about the nature of the texture being analysed. These attributes may partially explain its relatively good performance in OPT analysis.

The class balancing problem is important and the performance can be further improved if it is carefully managed. Using the approximation of non-linear kernels, especially χ^2 kernel, also improves the classification performance slightly when the dimensionality parameter is properly chosen.

In terms of computational complexity, encoding a window with RP and ISA methods is similar: the output window feature is a simple multiplication of the vectorised window with a matrix (either a random projection matrix[4] or a set of filters learned with ISA). The LBP operator[5] is more expensive as it involves a neighbourhood-pixel thresholding and a uniform pattern matching procedure. ISA requires a training phase in order to estimate filters from training patches. In our experiments learning a set of 300 filters[6] from a $81 \times 81 \times 81$ patch set required approximately 4 hours; encoding and

[4]The complexity of RP method can be further reduced by constructing sparse random projections with a simpler sampling distribution than a standard normal distribution [e.g., 1, 87]. We use the standard normal distribution for RP matrix because both methods achieve similar performance and a standard normal distribution is easier to implement.

[5]For LBP operator we use the C++ implementation by [160] (*URL:* http://www.cse.oulu.fi/CMV/Downloads/LBPSoftware).

[6]For ISA model estimation we use the Matlab implementation by [74] (*URL:* http://www.naturalimagestatistics.net).

classifying 200 test patches with size $81 \times 81 \times 81$ required approximately 30 seconds for the RP and ISA methods, and 4 minutes for the LBP method on a 3.4GHz Intel i7 CPU with 16Gb memory.

LGD was the class most readily distinguished. Unsurprisingly, being the intermediate class, HGD was the most often confused class. In terms of problem formulations, ordinal regression performance was slightly worse than multi-class classification in terms of mis-classification rate, absolute error rate and averaged F-measure (Table 5.2). However, ordinal regression makes less confusion between the ICA and LGD classes (Figure 5.3). The ordinal regression formulation is a better choice over one-vs-rest classification when the focus concentrates on minimising the risk of mis-classification between LGD and ICA. Ordinal regression is also simpler than multi-class classification in the sense that only one model is trained while in multi-class classification three models are trained. Dysplastic change is naturally a continuous phenomenon on which ordinal grading imposes artificial categories. The ordinal regression method works by first mapping samples onto the real line and it would be interesting to investigate using this continuous map in diagnostic histopathology.

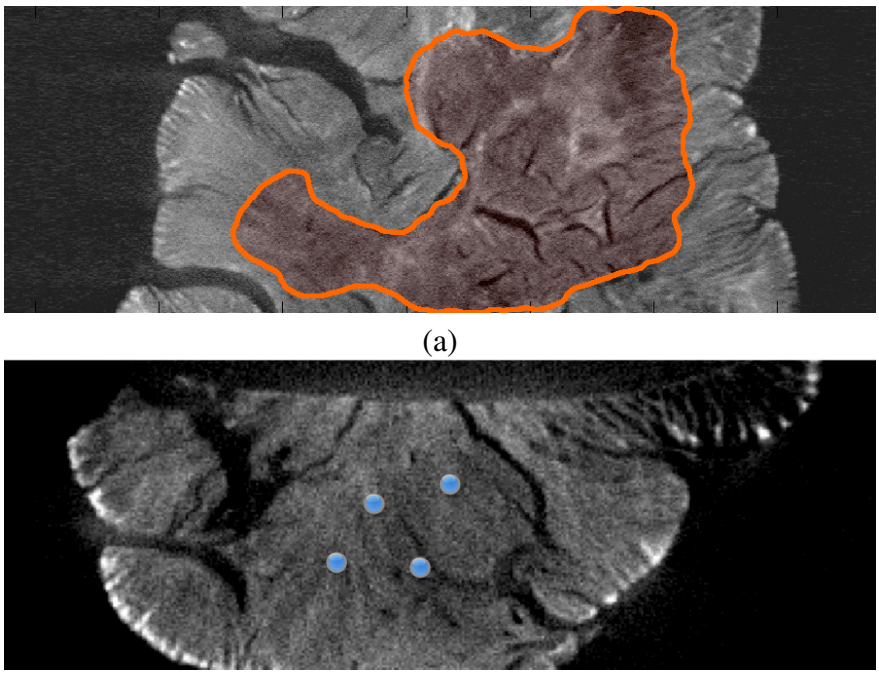
Chapter 6

Cancer detection with partial annotations

6.1 About this chapter

To model the underlying patterns of image regions, accurate annotations are necessary. However, the volumetric images of polyps are large (1024^3 voxels); while high resolution brings us considerable detail, difficulty arises in obtaining annotations. In OPT dataset, a polyp typically extends across $700 \sim 800$ slices and about 0.5 billion voxels in total in one OPT image. Fully delineating 3D regions slice by slice is tedious and time-consuming.

This chapter presents an alternative approach based on partial, sparse, incomplete annotations. A learning framework is proposed for partially annotated OPT image for the task of cancer detection in colorectal polyps. The focus of this chapter is the cancer detection task because it is one of the most important problems in medical image analysis. More specifically, the objective in this chapter is to discriminate between image patches in two settings: (1) invasive cancer (ICA) vs. low-grade dysplasia (LGD) and (2) ICA vs. the other classes (LGD and HGD).



(a)

(b)

Figure 6.1 OPT colorectal polyp images with (a) a region fully annotated and (b) some partial annotations.

Different forms of partial annotation can be appropriate for different image modalities and applications. In this chapter, we consider partial annotations consisting of just one click or a few clicks in the 3D polyp region of interest (as shown in Figure 6.1(b)) as an alternative to the stronger annotation shown in Figure 6.1(a). The annotation effort required is quite different. Our goal is to reduce the annotation efforts while achieving good classification performance. In addition, learning should scale well making it suitable for high-resolution volumetric images.

6.2 Related work

In Chapter 4 and 5, local features for patch and region classification of OPT images were compared. Here we focus on the model learning aspect of the task. Our method falls into the broad category of weakly supervised classification. At one extreme of this category, annotation is performed only at the image level, in which case multiple instance learning (MIL) has been adopted (we explored this case in Chapter 7). In

MIL, a sample image is labelled as positive if and only if at least one of the instances is classified as positive. Dundar et al. [41] proposed a large-margin approach for pathology slides. It shared some similarity to our work, however the prediction was at the image level. In Xu et al. [149] MIL was adapted to classify and segment histopathology images. Doyle et al. [40] applied active learning to detect cancer regions with histopathology annotations. Our approach is to leverage spatial annotation but to keep annotation simple, sparse and thus fast to perform.

6.3 Methods

In supervised classification, locations outside annotated regions are usually ignored during training because the corresponding class labels are considered unknown. However, for images annotated with a partial annotation protocol, the annotations carry information about the class membership at unannotated locations. We refer to 3D cubic regions as *patches*. Patches in the training set at locations with annotated (known) class labels are referred to as *reference patches*. Patches near to them (in terms of displacement or distance in feature space) are referred to as *candidate patches*. In this chapter, we consider an extreme form of partial annotation consisting of single point locations defined by mouse clicks. We introduce our definition of contextual relevance, based on which we then propose a ranking model for classification. Figure 6.2 illustrates the reference and candidate patches in feature classification. Intuitively the decision boundary was found so that the distances between reference patches and the decision boundary are large (classified with high confidence), while the distances between candidate patches and the boundary are small (classified with low confidence).

6.3.1 Labelling patches' confidence

First we assign confidence labels to candidate patches. Consider a reference patch $\mathbf{S}_r$ sampled at an annotated location $\mathbf{z}_r$ labelled as $y_r \in \{1, -1\}$. The patch $\mathbf{S}_k$ sampled at a nearby unannotated location $\mathbf{z}_k$ will have a lower confidence label y_k which can be set

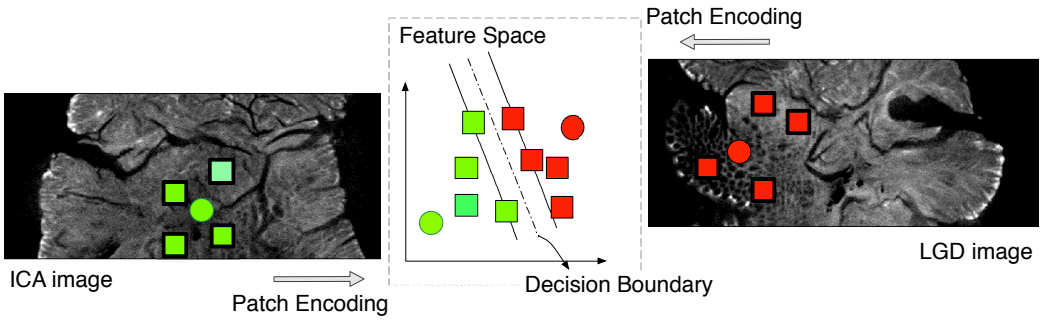


Figure 6.2 ICA-vs-LGD image patch classification with reference patches (circles) and candidate patches (rectangles).

to:

$$y_k = a(\mathbf{S}_k, \mathbf{S}_r)\, y_r, \tag{6.1}$$

where $a(\cdot,\cdot) \in (0, 1]$ is a measurement of affinity between two image patches. The absolute value of y_k can be viewed as a confidence measurement.

As patches sampled at locations near to each other usually belong to the same class, the reference patch of $\mathbf{S}_k$ can be set as the nearest annotated patch $\mathbf{S}_r$. Affinity $a(\cdot,\cdot)$ is defined as a Gaussian function with regard to spatial displacement of $\mathbf{S}_k$ and $\mathbf{S}_r$ in the image and a scaling parameter σ, i.e.,:

$$a(\mathbf{S}_k, \mathbf{S}_r) = \exp(-\frac{\|\mathbf{z}_k - \mathbf{z}_r\|^2}{\sigma^2}). \tag{6.2}$$

Another way to define $a(\cdot,\cdot)$ is to consider similarity in feature space. Assuming feature $\mathbf{x}_r$ is extracted from reference patch $\mathbf{S}_r$ and $\mathbf{x}_k$ from $\mathbf{S}_k$, we can define the feature-based affinity with a scaling parameter δ as:

$$a(\mathbf{S}_k, \mathbf{S}_r) = \exp\,(-\frac{\|\mathbf{x}_k - \mathbf{x}_r\|^2}{\delta^2}). \tag{6.3}$$

Note that $a(\cdot,\cdot)$ can be extended to use multiple reference patches. σ and δ are free parameters modelling how fast the confidence decreases when the distance from candidate to reference patch increase. Here we assign only one (the nearest) reference patch for each candidate patch.

6.3.2 Contextual relevance ranking model

Let $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \cdots, N$ denote a feature vector extracted from an image patch indexed by i. We assign the label $y_i \in \{1, -1\}$ if the i-th feature vector is from a reference patch; otherwise we set y_i according to formula (6.1). We form the ranking model by optimising a regularised margin-based problem:

$$\min_{\mathbf{w}, b} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{6.4}$$

$$\text{s.t.} \quad \frac{1}{y_i}(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall \mathbf{x}_i \in \mathbf{X}, \tag{6.5}$$

$$\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j \geq R_{ij}, \quad \forall \mathbf{x}_i \in \mathbf{X}^+, \mathbf{x}_j \in \mathbf{X}^-, \tag{6.6}$$

where $\mathbf{X}^+ = \{\mathbf{x}_k : 0 < y_k \leq 1\}$, $\mathbf{X}^- = \{\mathbf{x}_k : -1 \leq y_k < 0\}$ and $\mathbf{X} = \mathbf{X}^+ \cup \mathbf{X}^-$. The pairwise contextual relevance R_{ij} of two patches $\mathbf{S}_i$ and $\mathbf{S}_j$ is defined as:

$$R_{ij} = \frac{a(\mathbf{S}_i, \mathbf{S}_{ir})a(\mathbf{S}_j, \mathbf{S}_{jr})}{a(\mathbf{S}_i, \mathbf{S}_{ir}) + a(\mathbf{S}_j, \mathbf{S}_{jr})}, \tag{6.7}$$

where the patches $\mathbf{S}_{ir}$ and $\mathbf{S}_{jr}$ are the reference patches of $\mathbf{S}_i$ and $\mathbf{S}_j$ respectively. Constraints (6.5) are for all feature vectors in the training set. Note that in (6.5) features from candidate patches y_k are loosely constrained compared to their reference patches $y_r \in \{-1, 1\}$ because $|y_k| \in (0, 1)$. Constraints (6.6) rank a pair of patches from two images with regard to their contextual relevance. We argue that patches sampled nearer to annotated locations (in image or feature space) should be classified with a larger score, i.e., further away from decision boundaries. The constraints keep a large projected distance between any data point with high magnitude in y and the data points in the opposite class. In the case that pair $(\mathbf{x}_i, \mathbf{x}_j)$ is labelled with certainty, i.e., $y_i = \pm 1$ and $y_j = \pm 1$, the pairwise constraint (6.6) vanishes due to constraint (6.5). Figure 6.3 illustrates the geometric interpretation of this model.

Given a training set, the optimisation problem can be transformed into a dual form of $\mathbf{w}$ by constructing a new feature set with $\frac{(\mathbf{x}_i - \mathbf{x}_j)}{R_{ij}}$. Then this can be solved by any

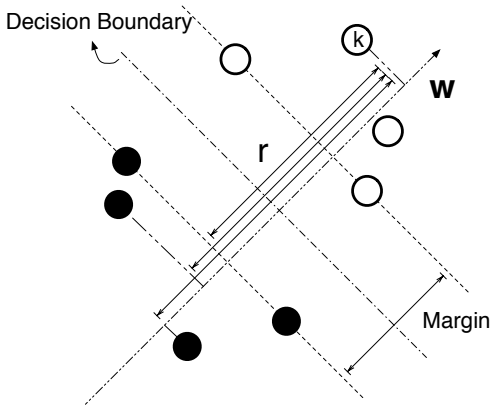


Figure 6.3 Geometric interpretation of the contextual relevance ranking model. $\mathbf{w}$ is the weight vector; point k is a feature vector extracted from a reference patch. r refers to the differences of ranking score between data point k and points in the other class (projected along the direction of $\mathbf{w}$). Constraints in formula (6.5) were designed for minimising the classification error; constraints in formula (6.6) were designed for optimising the ranking difference r.

SVM dual form solver, e.g., LIBSVM, SVM^{light}. However, this method is very slow and constructing feature set $\frac{(\mathbf{x}_i - \mathbf{x}_j)}{R_{ij}}$ across all the pairwise constraints is infeasible for our problem because the set of candidate patches is large. Here we tackle the primal form directly with a recently proposed efficient stochastic gradient method, SAG [124]. This method enables us to learn features online and with minimal storage cost.

To solve the optimisation problem we minimise function (6.4) while controlling constraint violations in (6.5) and (6.6). The risk function $J(\mathbf{w}, b)$ on training features $\{\mathbf{x}_i\}_{i=1}^N$ and labels $\{y_i\}_{i=1}^N$ can be written as:

$$J(\mathbf{w}, b) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{N^+N^-}\sum_{\mathbf{x}_i \in \mathbf{X}^+}\sum_{\mathbf{x}_j \in \mathbf{X}^-}(f_i + f_j + Cg_{ij}); \tag{6.8}$$

$$\text{where: } f_i = f(\frac{1}{y_i}(\mathbf{w}^T\mathbf{x}_i + b)), \quad f_j = f(\frac{1}{y_j}(\mathbf{w}^T\mathbf{x}_j + b)), \tag{6.9}$$

$$g_{ij} = g(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j - R_{ij}). \tag{6.10}$$

The loss term $f(\cdot)$ corresponding to constraints (6.5) is the squared hinge loss:

$$f(t) = \max(0,\ 1 - t)^2; \tag{6.11}$$

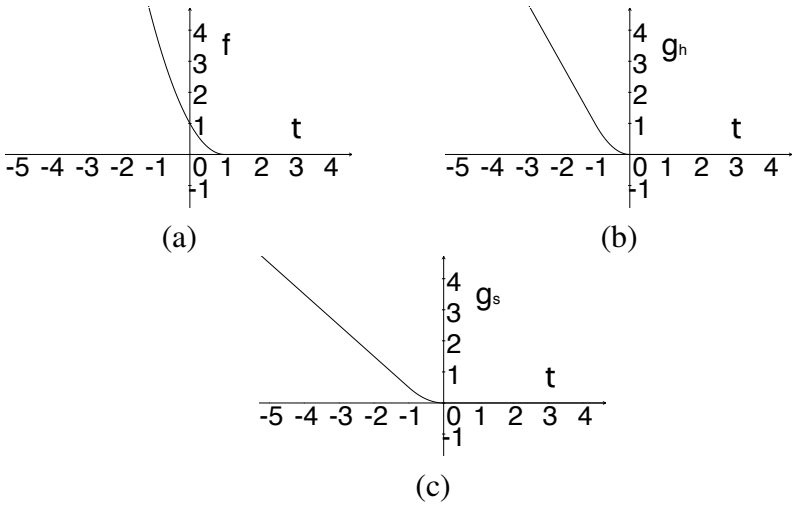


Figure 6.4 Demonstration of loss functions: (a) squared hinge loss $f(t)$, (b) Huber loss $g_h(t)$, and (c) smoothed hinge loss $g_s(t)$.

the loss term $g(\cdot)$ corresponding to constraints (6.6) can be a Huber loss function:

$$g_h(t) = \begin{cases} 0 & \text{if} \quad t \geq 0 \\ -2t - 1 & \text{if} \quad t < -1 \\ t^2 & \text{otherwise} \end{cases}, \tag{6.12}$$

or a smoothed hinge loss function:

$$g_s(t) = \begin{cases} 0 & \text{if} \quad t \geq 0 \\ -t - \frac{1}{2} & \text{if} \quad t < -1 \\ \frac{1}{2}t^2 & \text{otherwise} \end{cases}. \tag{6.13}$$

In risk function $J(\mathbf{w}, b)$, parameter λ is the regularisation strength; $C \geq 0$ controls the trade-off between classification errors and ranking errors; N^+ and N^- are the number of positive and negative samples respectively. Figure 6.4 illustrates the loss function used in $J(\mathbf{w}, b)$. The squared hinge loss is much more sensitive to outliers and large errors than the smoothed hinge and Huber loss. It is applied to patch classification to ensure the risk function is sensitive to every training label. The latter two functions are choices for the pairwise ranking errors. OPT images of colorectal polyps usually involve large intra-class variations so we expect the pairwise outliers would not dominate the

risk function. There are other loss functions that meet our requirements [122]. We chose these convex and smooth functions as they can be efficiently integrated into SAG.

To apply SAG methods for minimising $J(\mathbf{w}, b)$ iteratively, at each iteration $\mathbf{w}$ is updated with an average of the gradient of a randomly selected training pair $(\mathbf{x}_i, \mathbf{x}_j)$ and most recently computed gradients of the other training pairs. At the *(k+1)*th iteration the updating rule with a small step size α_k has the form:

$$\mathbf{w}^{k+1} = (1 - \alpha_k \lambda)\mathbf{w}^k - \frac{\alpha_k}{N^+ N^-} \sum_{\mathbf{x}_i \in \mathbf{X}^+} \sum_{\mathbf{x}_j \in \mathbf{X}^-} grad_{ij}^k, \tag{6.14}$$

where for the training pair (i^k, j^k), we set:

$$grad_{ij}^k = \begin{cases} f_i' + f_j' + Cg_{ij}' & \text{if}(i, j) = (i^k, j^k) \\ grad_{ij}^{k-1} & \text{otherwise} \end{cases}. \tag{6.15}$$

For the bias term b, we simply extend each feature vector with one bias component (from $\mathbf{x}$ to $[\mathbf{x}; b]$) in each iteration. This method has an exponential convergence rate and with a few implementation tricks (described in [124]) we reduce the storage cost to $O(N^+ N^-)$. This allows the method to scale to large datasets.

6.4 Evaluation

In this section we evaluated two aspects of the proposed model in terms of patch classification performance: (1) the ability to utilise both labelled and unlabelled patches (reference and candidate patches), compared with not using unlabelled patches, and using unlabelled patches naively (standard SVM); (2) the choice of loss function and affinity measurement. The experiments were conducted in two classification settings: (1) ICA-vs-LGD classification, and (2) ICA-vs-rest classification.

Experimental setup

In the experiments, we applied 10-fold cross-validation. 10 iterations of training and testing were performed such that within each iteration one fold was used as test set. The performance in terms of averaged area under the ROC curve (AAUC) values was averaged over the 10 iterations as the performance measure (as illustrated in Figure 5.2).

Test patches were randomly sampled from annotated regions in the test folds. In the training sets, the partial annotation process was simulated by randomly sampling point locations within the pathologist-annotated regions. These annotations were confirmed by the pathologist. Candidate patches were randomly sampled in the training set.

With reference and candidate patches, three types of models were trained:

- **T1 (standard supervised settings without candidate patches)**: training with only reference patches, using standard SVM (denoted as *SVM.REF*).
- **T2 (standard supervised approach with candidate patches)**: training with both reference and candidate patches, using standard SVM. Labels of candidate patches can be assigned with either feature-based or location-based affinity (formula (6.2) or (6.3)) (denoted as *SVM.ALL*).
- **T3 (proposed approach with candidate patches)**: using our proposed model with both reference and candidate patches. We evaluated four combinations of different loss functions (formula (6.12) and (6.13)) and affinities (formula (6.2) and (6.3)) (denoted as *PROP.HUBER* and *PROP.SQU*).

Method T2 used candidate patches in a standard SVM by assuming their labels are binary. The binary labels were determined by their reference patches according to formula (6.2) or formula (6.3). Method T3 considered the confidences of candidate patches using the proposed contextual ranking framework.

For feature extraction we used bag-of-words encoding with random projection since this achieved high classification accuracies in Chapter 5. The dimensionality of each

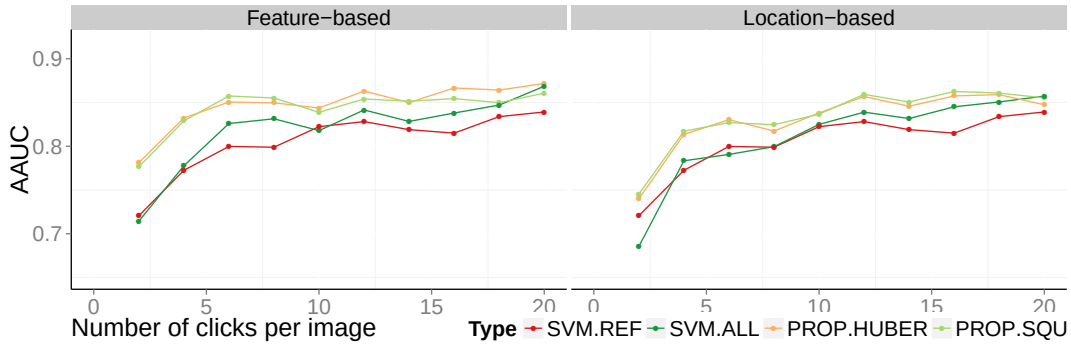
feature vector was 200. Each feature was normalised to zero mean and unit variance. In all standard SVM evaluations, we used the LIBLINEAR [44] solver that solves the ℓ_2 regularised squared loss primal problem (with regularisation parameter searched from 10^{-7} to 10^{7} and $eps = 0.01$). In our proposed method, C searched from 10^{-10} to 10^{-5}, $\lambda = \frac{1}{N^+N^-}$, $b = 0$, and the stochastic gradient step size was set to 0.004. The scaling factors were estimated from standard deviation of all distances ($\|\mathbf{z}_k - \mathbf{z}_r\|$ or $\|\mathbf{x}_k - \mathbf{x}_r\|$) between reference patches and candidate patches in the training set ($\sigma = 158.1$ in formula (6.2), $\delta = 7071.1$ in formula (6.3)).

6.4.1 Cancer-vs-LGD classification

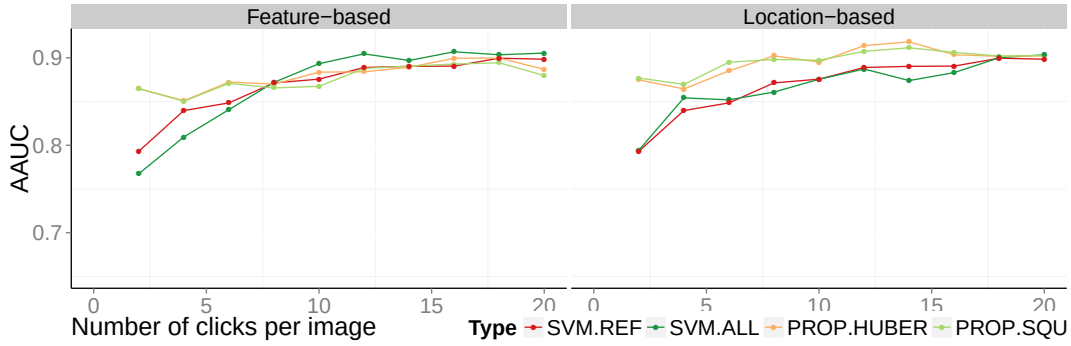
As a preliminary experiment, the proposed method was initially validated with ICA-vs-LGD classification using OPT images from 59 patients (30 LGD and 29 ICA). We started with a training set with only 2 reference patches sampled at the click locations and 40 candidate patches sampled outside the annotated regions for each training image — the candidate patches were sampled so that we have no knowledge of the ground truth labels of these patches.

The models were learned using T1, T2 and T3 methods respectively and the classification performance was evaluated on the test patches. Then we added more reference patches and their associated candidate patches. At each iteration 2 reference and about 15 candidate patches per image were added. Such iterations were repeated 20 times till there were about $1,500$ reference patches in the training set. At the final iteration the number of training patches was about $10,000$.

Figure 6.5 shows AAUC values depending on the number of reference patches per training image at patch size $21 \times 21 \times 21$ and $81 \times 81 \times 81$ respectively. We list the AAUC values depending on number of reference patches per image in Table 6.1. In Figure 6.5, AAUC is generally higher at patch size $81 \times 81 \times 81$ than $21 \times 21 \times 21$. With more than 10 reference patches per image (530 reference patches, about 3,000 training patches in total) the classification performances of all the methods saturated. With both



(a) at patch size $21 \times 21 \times 21$.



(b) at patch size $81 \times 81 \times 81$.

Figure 6.5 AAUC values (ICA-vs-LGD) depending on number of reference patches with location-based and feature-based affinity measurements.

location-based and feature-based affinity the SVM.ALL method showed the same or slightly higher AAUCs than the SVM.REF method. This indicates that simply feeding uncertain patches to standard SVM does little to help patch classification performance. The information presented in uncertain patches was not utilised effectively by standard SVM. The proposed methods performed relatively well with small training sets indicating that they were making effective use of the unannotated patches. AAUCs of all methods converged to similar values when number of reference patches reaches 20 per image.

For both affinity-based experiments, the proposed models with Huber loss and smoothed hinge loss showed almost the same AAUCs. However, our grid search of parameters showed that the best parameters C are quite different ($C = 10^{-4}$ for Huber loss and $C = 10^{-2}$ for smoothed hinge loss).

Table 6.1 Cancer-vs-LGD classification performance comparison between standard SVM and proposed model. AAUC values ± standard errors depending on the number of reference patches per image.

((a)) Feature-based Affinity, at patch size $21 \times 21 \times 21$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.72 ± 0.029	0.77 ± 0.014	0.80 ± 0.023	0.83 ± 0.021
SVM.ALL	0.71 ± 0.022	0.78 ± 0.030	0.83 ± 0.022	0.84 ± 0.023
PROP.HUBER	**0.78 ± 0.022**	0.83 ± 0.020	0.85 ± 0.019	**0.86 ± 0.016**
PROP.SQU	0.78 ± 0.023	**0.83 ± 0.018**	**0.86 ± 0.018**	0.85 ± 0.019

((b)) Location-based Affinity, at patch size $21 \times 21 \times 21$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.72 ± 0.029	0.77 ± 0.014	0.80 ± 0.023	0.83 ± 0.021
SVM.ALL	0.69 ± 0.022	0.78 ± 0.030	0.80 ± 0.022	0.84 ± 0.023
PROP.HUBER	0.74 ± 0.022	0.81 ± 0.020	0.82 ± 0.019	**0.86 ± 0.016**
PROP.SQU	**0.75 ± 0.023**	**0.82 ± 0.018**	**0.82 ± 0.018**	0.86 ± 0.019

((c)) Feature-based Affinity, at patch size $81 \times 81 \times 81$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.79 ± 0.028	0.84 ± 0.037	0.87 ± 0.027	0.89 ± 0.026
SVM.ALL	0.77 ± 0.040	0.81 ± 0.034	**0.87 ± 0.023**	**0.90 ± 0.014**
PROP.HUBER	0.86 ± 0.026	**0.85 ± 0.021**	0.87 ± 0.024	0.88 ± 0.025
PROP.SQU	**0.86 ± 0.025**	0.85 ± 0.022	0.87 ± 0.025	0.89 ± 0.026

((d)) Location-based Affinity, at patch size $81 \times 81 \times 81$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.79 ± 0.027	0.84 ± 0.037	0.87 ± 0.027	0.89 ± 0.026
SVM.ALL	0.79 ± 0.040	0.85 ± 0.028	0.86 ± 0.032	0.89 ± 0.022
PROP.HUBER	0.87 ± 0.021	0.86 ± 0.026	**0.90 ± 0.022**	**0.91 ± 0.017**
PROP.SQU	**0.88 ± 0.021**	**0.87 ± 0.024**	0.89 ± 0.022	0.90 ± 0.019

6.4.2 Cancer-vs-rest classification

In addition to the ICA-vs-LGD classification, the experiments were extended to include the high-grade dysplasia (HGD) class. 90 images were used with 30 from each of the three classes (LGD, HGD and ICA). The aim of this experiment was a fair comparison using the same setting described in Chapter 5, except that the classification models were trained with partial annotations in this experiment.

Figure 6.6 AAUC values (ICA-vs-rest) depending on number of reference patches with location-based and feature-based affinity measurements.

Table 6.2 Cancer-vs-rest classification performance comparison between standard SVM and proposed model. AAUC values ± standard errors depending on the number of reference patches per image.

((a)) Feature-based Affinity, at patch size $81 \times 81 \times 81$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.81 ± 0.032	0.81 ± 0.032	0.82 ± 0.030	0.83 ± 0.030
SVM.ALL	0.68 ± 0.021	0.72 ± 0.025	0.80 ± 0.029	0.79 ± 0.032
PROP.HUBER	**0.81 ± 0.025**	0.81 ± 0.020	0.83 ± 0.023	0.83 ± 0.020
PROP.SQU	0.80 ± 0.025	**0.82 ± 0.021**	**0.83 ± 0.022**	**0.84 ± 0.020**

((b)) Location-based Affinity, at patch size $81 \times 81 \times 81$.

# Clicks Per Image	2	4	8	12
SVM.REF	0.81 ± 0.032	0.81 ± 0.032	0.82 ± 0.030	0.83 ± 0.030
SVM.ALL	0.71 ± 0.036	0.73 ± 0.036	0.81 ± 0.031	0.83 ± 0.027
PROP.HUBER	**0.81 ± 0.026**	0.82 ± 0.026	**0.85 ± 0.018**	0.83 ± 0.021
PROP.SQU	**0.81 ± 0.026**	**0.82 ± 0.025**	0.85 ± 0.019	**0.83 ± 0.020**

A binary ranking model was evaluated by treating ICA as the positive training samples and both LGD and HGD as the negative training samples. We followed the same procedure of adding reference patches into the training process as in the previous section. Unlike the previous experiments, the candidate patches were sampled randomly from the training images without using the region annotations. This is a more realistic approach to obtain the candidate patches. Figure 6.6 shows AAUC values depending on the number of reference patches per training image at patch size $81 \times 81 \times 81$. We list the AAUC values depending on the number of reference patches per image in Table 6.2.

We observed similar trends of classification performance to that in Section 6.4.1. In terms of AAUC, our method is slightly better than training SVM with only reference patches (the SVM.REF method). However, it is often much better than simply adding uncertain patches into the training set (the SVM.ALL method). The difference between different loss functions, i.e., between the PROP.HUBER and the PROP.SQU method was very small. The overall highest AAUC achieved, by using the location-based affinity measure was 0.85 ± 0.19 at 8 clicks per image. Not surprisingly, the AAUC is lower than that of using all the fully labelled image patches in supervised settings described in Section 5.6, where AAUC 0.876 was achieved for cancer-vs-rest patch classification.

6.5 Summary

A learning model was proposed for partially annotated images. The experiments conducted on cancer-vs-LGD and cancer-vs-rest classifications showed that it is able to robustly learn from patches with uncertain labels, achieving high classification performances while reducing the annotation effort. At the same time, the proposed model can be efficiently evaluated with only $O(N^+N^-)$ in storage cost. Therefore it is suitable for high-resolution, volumetric datasets.

Chapter 7

Cancer detection with image-level annotations

7.1 About this chapter

The previous chapter described training a cancer detector with click annotations; this chapter presents training with image-level annotations using a multiple instance learning (MIL) framework. MIL has recently been applied to histopathology image analysis for both segmentation and classification tasks [80, 150]. MIL methods can potentially infer cancerous regions with image-level annotation, i.e., binary labels indicating whether cancer is present in the image. The MIL formulation is attractive as it does not require the effort of manually delineating image regions.

The general MIL inference rules are defined in the context of binary classification: a bag of instances is positive if at least one instance in the bag is positive, negative if all of the instances in the bag are negative. A common implementation of the rules in image classification treats each image as a bag, and regions in an image as instances. In terms of histopathology image analysis, an example application is to label an image as *cancer* if cancer is present in at least one region of the image, and as *non-cancer* otherwise.

In this chapter, following the MIL setting, we propose a novel tree boosting algorithm for training a cancer detector. Our algorithm extends multiple instance boosting (MILBoosting) [154] by boosting regularised trees with instance-to-prototype distances as features. The discriminative prototypes in our algorithm are searched by solving a submodular set cover problem. Our approach is validated with 2-D OPT image frames, 3-D OPT volumes, as well as a public dataset of breast cancer tissue microarray (TMA) images.

7.2 Related work

Although MIL has been extensively studied since [102] and there exists a large literature (for a general review of MIL, see [7]), it was only recently applied to histopathology image analysis. Here we give a brief review of some of the most relevant work.

Zhao et al. [159] applied multiple-instance learning via embedded instance selection (MILES) [25] for 10 category histopathology image classification. Xu et al. [150] extended MILBoosting [154] to simultaneously detect and cluster multiple types of tissue region in TMA images. Kandemir et al. [79] evaluated MIL formulations on diagnosis of Barrett's cancer with H&E images. Xu et al. [148] used MIL to classify colon cancer histopathology images with features extracted from convolutional neural networks.

Selecting instances as prototypes for bag classification was used previously with bags represented in terms of distances to prototypes [25, 51]. Our work extends MILBoosting to select prototypes with instance-to-prototype distances. We search a set of positive instance prototypes that is both discriminative and covers multiple modes of the appearance distribution. Instance-to-prototype distances are considered as features. A regularised regression tree boosting method is proposed to further select and combine the features.

Prototypes should satisfy three criteria; they should be (1) relevant: present in many positive images, (2) discriminative: dissimilar to negative instances, and (3) complementary: covering multiple types of positive instances. Song et al. [134] formalised these intuitions as a submodular set cover problem solved by a greedy algorithm. The set of prototypes was used as an initial training set for latent SVM. In this chapter, we adopt the 'discriminativeness' of each prototype as a regularisation strength in the MILBoosting framework.

7.3 Methods

7.3.1 Notation

Here we introduce the notation adopted throughout the chapter. We denote $\mathbf{x}_{ij} \in R^d$ as a d-dimensional feature representation of an instance (patch). Index ij represents the j^{th} instance in the i^{th} bag (image). $y_i \in \{0, 1\}$ represents the label of the bag, where 0 denotes non-cancer and 1 denotes cancer. The k^{th} prototype $\mathbf{p}_k \in R^d$ is an instance selected from the training instance set. In the following sections we introduce the two steps of our proposed method: searching for a set of discriminative prototypes and learning cancer detectors.

7.3.2 Discriminative prototypes

The discriminativeness of a prototype $g(\mathbf{p}_k)$ can be estimated as follows [134]: first find the m nearest neighbours of $\mathbf{p}_k$ from the set of training instances $\{\mathbf{x}_{ij}\}$; then, counting the number of neighbours from the positive bags (denoted as m_{pos}), the ratio m_{pos}/m can be a measurement of discriminativeness. Greedy search for a set of prototypes starts with an empty set of prototypes and a candidate set comprising all training instances. The most discriminative instance from the candidate set is then added to the prototype set, at the same time removing the prototype's m nearest neighbours from the candidate

set. This is repeated until the candidate set is small enough (e.g., only 10% of initial candidates remain).

Song et al. [134] used the prototypes as an initialisation set for training latent SVM. We propose to combine the set of prototypes in a boosting framework. We further select prototype subsets and simultaneously learn an instance classifier by boosting regularised trees where we utilise $g(\mathbf{p}_k)$ as the regularisation strength.

7.3.3 Boosting with regularised regression trees

In MILBoosting, the instance classifier $F(\mathbf{x}_{ij})$ is formulated as a linear combination of T weak learners, i.e., $F(\mathbf{x}_{ij}) = \sum_{t=1}^{T} \alpha_t f_t(\mathbf{x}_{ij})$, where $f_t(\mathbf{x}_{ij})$ gives a score to each instance $\mathbf{x}_{ij}$; α_t is the weight of f_t. The probabilities that cancer is present in an instance P_{ij} and in a bag P_i are respectively modelled as

$$P_{ij} = \frac{1}{1 + \exp(-F(\mathbf{x}_{ij}))} \qquad \text{and} \qquad P_i = 1 - \prod_j (1 - P_{ij}). \tag{7.1}$$

The instance classifier can be estimated by minimising the negative log-likelihood L of the bag labels:

$$L(y_i, F(\mathbf{x}_{ij})) = -\log \prod_i P_i^{y_i} (1 - P_i)^{(1-y_i)}. \tag{7.2}$$

Using the gradient boosting framework [49], L can be optimised by iteratively fitting weak learners f_t and optimising coefficients α_t. We adopt J-terminal regression trees as weak learners with a boosting shrinkage parameter ν [67]. However, instead of fitting regression trees to the feature set $\{\mathbf{x}_{ij}\}$, we first represent each instance in terms of distances to prototypes, i.e.,

$$\hat{\mathbf{x}}_{ij} = [d(\mathbf{x}_{ij}, \mathbf{p}_1), \ldots, d(\mathbf{x}_{ij}, \mathbf{p}_k)], \tag{7.3}$$

where $d(.,.)$ is a distance measure, e.g., ℓ_2-distance. Regression trees are then constructed on the new feature set $\{\hat{\mathbf{x}}_{ij}\}$. Each of the regression trees partitions the feature

space into disjoint regions. The best variables to split and the optimal thresholds of the tree are searched by maximising information gain. In $\{\hat{\mathbf{x}}_{ij}\}$ each variable is associated with a prototype $\mathbf{p}_k$. Our method encourages trees constructed on $\{\hat{\mathbf{x}}_{ij}\}$ to split at those variables that are associated with large $g(\mathbf{p}_k)$. The motivation is to further select prototypes so that regression trees split at a few very discriminative prototypes, instead of splitting at many non-informative prototypes which could result in poor generalisation.

We introduce regularisation to properly control the variable to split in the tree construction process. The method we adopted is guided regularisation for tree construction [36]. It was first proposed as a feature selection technique integrated in a random forest classifier; the selection of variables was guided by pre-computing variable importance from a preliminary random forest training. Here we combine the regularisation method with boosting trees. We utilise the discriminativeness $g(\mathbf{p}_k)$ as the regularisation strength instead of a preliminary random forest training.

Specifically given a set of K prototypes, we normalise $g(\mathbf{p}_k)$ as

$$\hat{g}(\mathbf{p}_k) = \frac{g(\mathbf{p}_k)}{\max_{k=1}^{K} g(\mathbf{p}_k)}; \tag{7.4}$$

when the tree chooses to split on the k^{th} feature of $\{\hat{\mathbf{x}}_{ij}\}$, the information gain is regularised by a function of $\hat{g}(\mathbf{p}_k)$:

$$\mathrm{G_R}(k) = ((1-\lambda)\gamma + \lambda\hat{g}(\mathbf{p}_k))\mathrm{Gain}(k), \tag{7.5}$$

where $\lambda \in [0, 1]$ is a free parameter to control the overall regularisation; $\gamma \in [0, 1]$ is a base regularisation coefficient. We calculate $\mathrm{Gain}(k)$ as the reduction of variance at all leaf nodes when splitting at the k^{th} feature. The regularised regression tree can directly utilise discriminativeness to control the regularisation strength via Formula (7.5). The tree-based feature selection can capture non-linear variable interactions if $J > 2$. We set $J = 4$, $\gamma = 1$, and grid search for λ in our experiments. Algorithm 1 summarises the proposed procedure.

Algorithm 1 Summary of the proposed algorithm

procedure BOOSTING PROTOTYPES($\{\mathbf{x}_{ij}\}, \{y_i\}, \nu, J, T$)
  $\{\mathbf{p}_k\} \leftarrow$ Greedy search for prototypes (Section 7.3.2)
  $\{\hat{\mathbf{x}}_{ij}\} \leftarrow$ Transform $\{\mathbf{x}_{ij}\}$ with $\{\mathbf{p}_k\}$ (Formula (7.3))
  for $t \leftarrow 1 \cdots T$ **do**
    for all i,j **do**
      Compute $r_{ij}^t \leftarrow -\frac{\partial L}{\partial f}\Big|_{f=f_{t-1}}$
    end for
    Fit a J-terminal regression tree f_t to r_{ij}^t (regularised by Formula (7.5))
    Line search: $\alpha_t \leftarrow \arg\min_\alpha L(y_i, F_{t-1}(\hat{\mathbf{x}}_{ij}) + \alpha f_t)$
    Update classifier: $F_t \leftarrow F_{t-1} + \nu\alpha_t f_t$ (shrinkage ν is a fixed parameter)
  end for
  return instance classifier F_T
end procedure

7.4 Evaluation

We evaluated the proposed method on cancer detection at (1) image level (predicting the presence of cancer in an unseen image) and (2) region level (localising the cancer region in an image). Two datasets were used in the experiments: a breast cancer TMA dataset and the colorectal polyps OPT dataset. For the OPT dataset, we evaluated MIL in two annotation configurations: the first is treating 2-D slice of OPT as bag, i.e., in the training stage, slice is manually labelled as either cancer or non-cancer. The second is treating sub-volumes of OPT as bags, i.e., label provided in the training stage only consists of whether cancer presented in a sub-volume of OPT polyp.

7.4.1 Experiments with breast cancer TMA images

The TMA dataset consists of 58 TMA breast cancer images stained with H&E. 26 images are diagnosed as malignant, 32 as benign. For a fair comparison we used the feature sets made publicly available[1] by Kandemir et al. [80]. Each image was divided into 49 equally-sized instances. Each instance was further encoded with a 708-dimensional feature vector. The feature vector composed of SIFT descriptors, local binary patterns, colour histograms, as well as cell-level morphological features.

[1]Link: http://www.miproblems.org/datasets/ucsb-breast/

Table 7.1 Cancer detection performance at image level measured with AAUC.

Method	GPMIL [80]	RGPMIL [80]	Proposed
AUC	0.86	0.90	0.93

Since the location information of each instance is not available in the feature set, we focus on image-level performance evaluation.

We follow the 4-fold cross validation protocol used in [80]. For the proposed method we first applied the set cover search with $m = 20$. This usually selects 100 to 200 positive prototypes from a total of $1,500$ instances. We set shrinkage parameter ν to 0.05, and the maximum number of iterations T to 300. The regularisation parameter λ was searched in the value set $\{0.1, 0.2, 0.3, \ldots, 1.0\}$ with a 10-fold cross validation on the training folds. Averaged area under the ROC curve (AAUC) was computed as the image classification performance measure (Table 7.1). The standard error of our method was 0.04. The equal Error Rate was 0.16 ± 0.03. Note that Relational Gaussian Process MIL (RGPMIL) was designed for TMA images by explicitly modelling cells with a graph. Both GPMIL and RGPMIL outperformed widely-used MIL methods including EMDD [158], MILBoosting [154] and MI-SVM [11]. As shown in Table 7.1 our method achieved better image-level performance than the top-ranked methods.

7.4.2 Experiments with 2-D OPT slice as bag

2-D slice dataset

To evaluate the feasibility of MIL setting for OPT image classification, we first conducted preliminary experiments by applying the proposed methods on 2-D slices of OPT image. We evaluated both image- and instance-level cancer detection performance on 60 OPT images (30 ICA and 30 LGD). The 2-D dataset consists of 200 2-D slices, with 100 slices randomly selected from ICA polyps and 100 from LGD. Figure 7.1 illustrates the experimental settings. Figure 7.3 shows some of the cancer slices and their annotations.

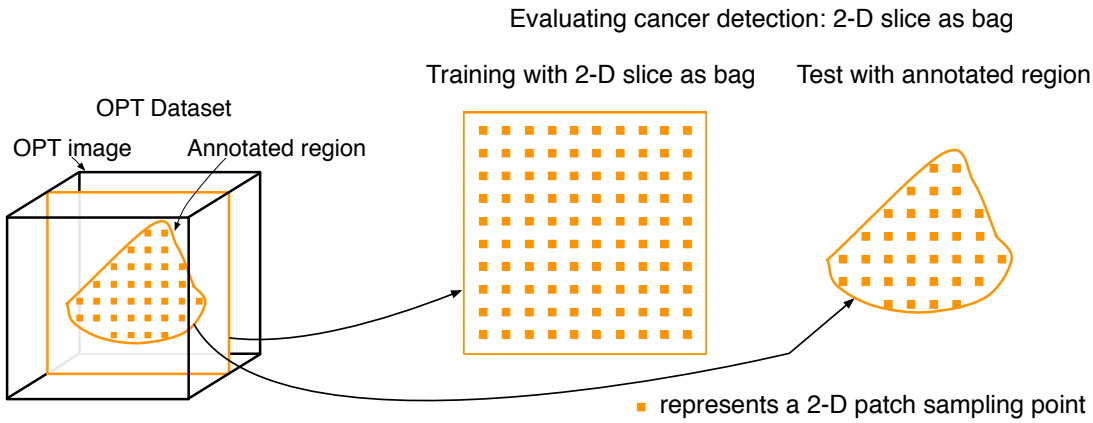


Figure 7.1 Illustration of evaluating with 2-D OPT slice as bag.

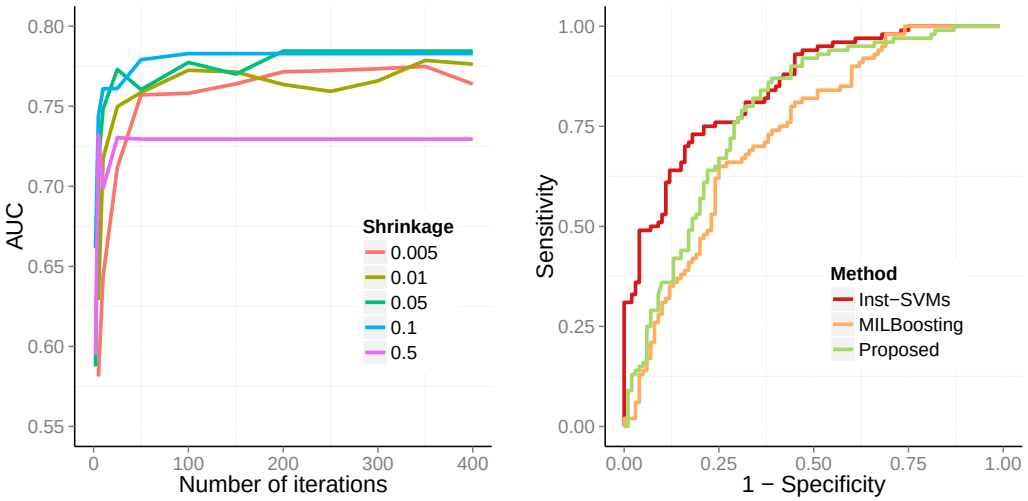
The pathologist was only asked to delineate major cancer regions with relatively high confidence rather than exhaustively trace all the cancer locations. Part of our motivation for applying MIL is that complete region-level annotations are difficult to obtain and validate. As a result, in ICA images, instance labels outside annotated regions were unknown. In the training stage, since the instance labels are not required in training MIL classifiers, the region annotations are not used. In the test stage, we report instance-level test results based on the classifier output over all instances in LGD images, and all instances that have at least 50% overlap with ICA annotations.

Experimental protocol

We treated each slice as a bag and densely extracted patches as instances. The size of each instance was 48×48 pixels. The sampling step size was 24 pixels in the training stage, and 12 in the test stage in both horizontal and vertical directions. We combined local binary patterns, SIFT features, and intensity histograms as instance features. The set cover search parameter was $m = 10$. Three-fold cross validation was conducted with the proposed method using the same grid search of parameters described in Section 7.4.1. Figure 7.2(a) shows the image-level AUC of the proposed method plotted against the parameters T and ν on a validation set. The performance in terms of AUC is not very sensitive to T and ν. Choosing a large T and a small ν tends to give a high AUC. We set ν to 0.05, and T to 300.

We also implemented MILBoosting as a baseline. Differences between the proposed method and MILBoosting are that the latter fits regression trees directly to $\{\mathbf{x}_{ij}\}$ without transformation and regularisation. In addition to the MIL methods we trained instance-level support vector machines (Inst-SVM) in a *fully supervised* setting as a comparison. In training Inst-SVM, instances with at least 50% overlap with the annotations were treated as cancer; the instances from LGD images were treated as non-cancer.

Results



(a) (b)

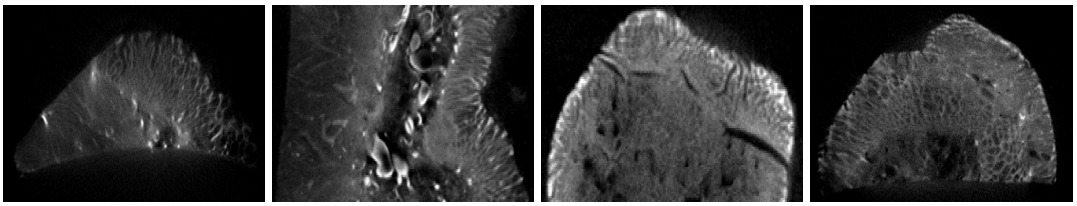
Figure 7.2 Cancer detection at image-level on 2-D slice dataset. (a) AUC of the proposed method against number of iterations T and shrinkage parameter ν, (b) ROC curves for the three methods compared.

Table 7.2 Cancer detection at image-level and instance-level (with standard errors).

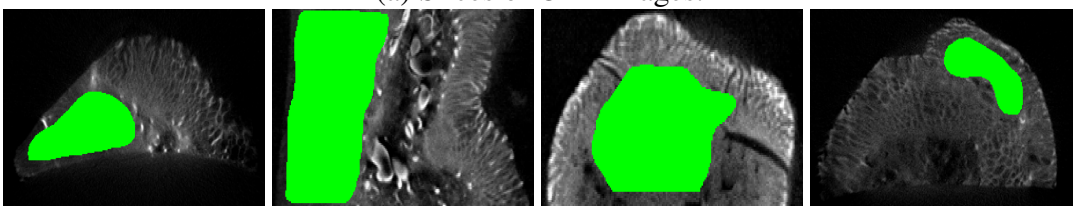
Method	MILBoosting	Proposed	Inst-SVMs
AUC (image-level)	0.74 ± 0.04	0.79 ± 0.01	0.85 ± 0.03
F-measure (instance-level)	0.41 ± 0.01	0.45 ± 0.03	0.53 ± 0.05

Table 7.2 compares image- and instance-level performance in terms of AUC and F-measure respectively. Figure 7.2(b) shows ROC curves at image-level. At image-level, Inst-SVM score was calculated as the maximum score in the image. At instance-level score thresholds of each method were searched on the training set by maximising

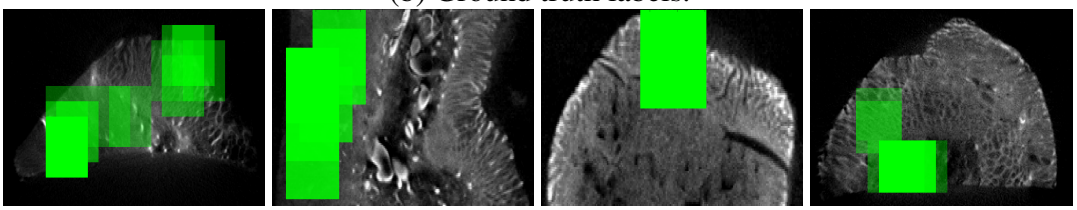
training F-measures. Our method outperformed MILBoosting in both image- and instance-level classification. However it was worse than fully supervised classifications, as would be expected. Figure 7.3 shows a few cancer detection examples.



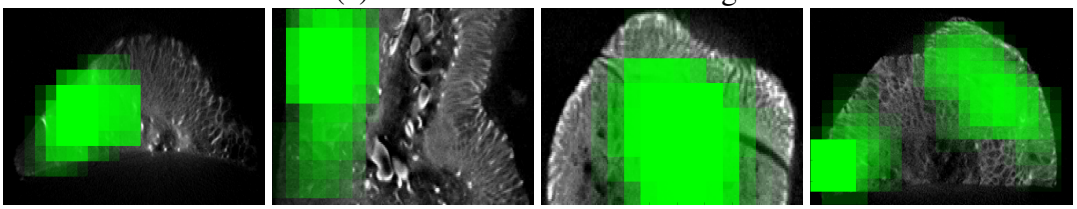
(a) Slices of OPT images.



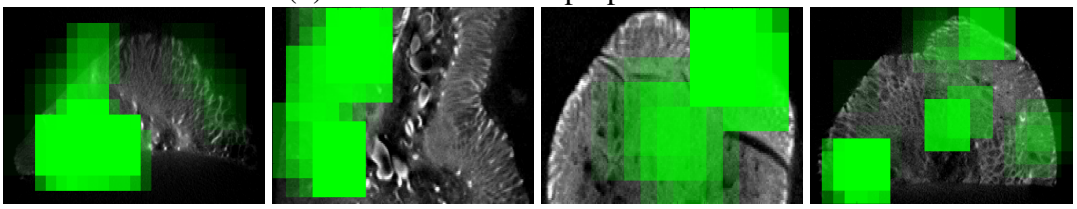
(b) Ground truth labels.



(c) Predictions of MILBoosting.



(d) Predictions of the proposed method.



(e) Predictions with fully supervised SVM.

Figure 7.3 Instance-level annotations and predictions. Green patches in third to fifth rows indicate scores of the instances are greater than the learned threshold. Instances with higher scores were mapped to higher opacity values.

7.4.3 Experiments with sub-volume of OPT polyp as bag

Sub-volume dataset

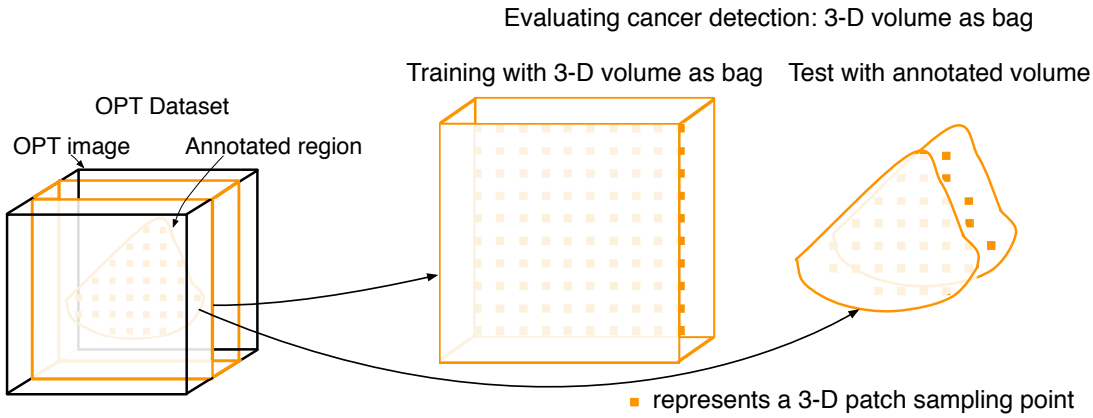


Figure 7.4 Illustration of evaluating with 3-D OPT slice as bag.

We further evaluated the proposed method on the entire dataset of 90 OPT images. Instead of using 2-D slices, our sub-volume dataset for MIL consists of 300 sub-volumes, with 100 volumes selected from ICA polyps and 200 from LGD and HGD. The 200 sub-volumes were treated as non-cancer in this experiment. The size of the sub-volume was $1024 \times 1024 \times 81$. The third dimension was set to size 81 in order to use the same parameter settings and have a fair comparison to the supervised classification conducted in Section 5.4. Figure 7.4 illustrates the experimental settings.

We cropped the sub-volume so that the black borders outside the polyps were removed and treated each cropped sub-volume as a bag. 3-D image patches densely sampled from the sub-volume were treated as instances. The size of each instance was $81 \times 81 \times 81$ voxels. The sampling step size was 40×40 voxels in the training stage. Each instance was encoded with bag-of-words framework using random projection features (described in Chapter 4).

Different from the previous Section 7.4.2, we follow exactly the same tenfold cross-validation scheme that was used in the supervised patch classification (Section 5.4) and in the experiments of learning with partial annotations (Section 6.4.2). In the training fold, MIL settings was applied to the sub-volume to train 3-D patch classifier, while

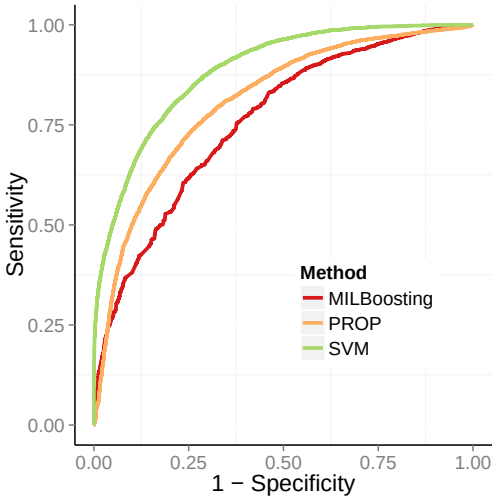


Figure 7.5 ROC curves for sub-volume cancer detection.

in the testing stage the classifier is tested on the delineated regions that were used in previous experiments in Section 5.4. The ROC curves with the method described in Section 7.4.2 were reported in Figure 7.5. The average AUC value was 0.81, 0.88, 0.76 for the proposed method, instance-level SVM, and MILBoosting [154].

Table 7.3 Comparison of patch classification performance measured with AAUC values ± standard errors.

Type of annotation	AAUC
Region annotation	0.88 ± 0.020
Click annotation	0.85 ± 0.019
Image-level annotation	0.81 ± 0.020

The same cross-validation scheme and test patch set were used for training with region annotations (Section 5.4), click annotations (Section 6.4.2), and image-level annotations (this section). The classification performances are compared in Table 7.3. The cancer-vs-rest patch classification performance increases as the level of details of the annotation increases.

7.5 Summary

This chapter introduced a novel multiple instance learning algorithm by combining discriminative prototype search with boosting of regularised regression trees. Compared to the work presented in Chapter 6, the requirement of manual annotations was further reduced to only image level. In the patch-level cancer detector training stage, only a binary training label was provided in each image that indicates whether cancerous regions are present in the image. The method is empirically studied for the task of cancer detection in OPT images. Initially, the method was validated with 200 slices sampled from the OPT dataset. Then the experiments were extended to the full dataset. In the training stage, binary labels were provided only at the sub-volume level. Additionally, the proposed methods were validated on a public breast cancer TMA dataset. All these experiments showed that the proposed method can achieve more accurate results in both image- and instance-level classification than the state-of-the-art methods.

Chapter 8

Conclusions

8.1 Summary of contributions

The aim of this research was to investigate automated methods for histological analysis of colorectal polyps in optical projection tomography. It is the first study on this topic. Efficient and effective feature extraction and classification methods, as well as the use of annotations from domain experts when training the systems, were presented.

Chapters 4 and 5 investigated the power of 3-D texture features to discriminate diagnostic levels of dysplastic change from OPT images, specifically, low-grade dysplasia, high-grade dysplasia, and invasive cancer. A patch-based recognition system was evaluated in both multi-class classification and ordinal regression formulations on a 90 polyp dataset. 3-D texture representations computed with a hand-crafted feature extractor, random projection, and unsupervised image filter learning were compared using the bag-of-words framework. The classification performance was measured in terms of error rates, F-measures, and ROC surfaces. Results demonstrated that randomly projected features were the best among the three texture representations. Discrimination was improved by carefully manipulating various important aspects of the system, including class balancing, output calibration and approximation of non-linear kernels. This work was published at the medical image understanding and analysis

(MIUA) conference 2013 [88] and the international symposium on biomedical imaging (ISBI) 2013 [90].

Annotations delineating regions of interest can provide valuable information for training medical image classification and segmentation methods. However the process of obtaining annotations is tedious and time-consuming, especially for high-resolution volumetric images. In Chapter 6, a novel learning framework to reduce the requirement of manual annotations while achieving competitive classification performance was presented. With images annotated with a few clicks, an image patch-based ranking model was developed to utilise the contextual information near the clicked locations. The results show that the proposed method can robustly infer patterns from partially annotated images with low computational cost. This is joint work with Dr. Wei-shi Zheng (Sun Yat-sen University) and has been published at the international conference on medical image computing and computer-assisted intervention (MICCAI) 2013 [91].

Learning cancer detectors using only image-level annotations is a very attractive yet challenging problem. Since it does not require the effort of manually delineating cancer regions, in practice it is more applicable than using the supervised learning setting. In Chapter 7, a novel multiple instance learning algorithm for cancer detection in colorectal polyp images was presented. With images labelled at slice level or sub-volume level, we first searched a set of image patch level prototypes by solving a submodular set cover problem. Regularised regression trees were then constructed and combined on the set of prototypes using a multiple instance boosting framework. The method compared favourably with competing methods in experiments on OPT images as well as on a public breast cancer tissue microarray dataset. This work has been accepted by the International Conference on medical image computing and computer-assisted intervention (MICCAI) 2015 [89].

In Chapter 5 we showed that by training the classification system with the regions annotated by the pathologist, an AAUC value of 0.88 was achieved for the task of cancer-vs-rest image patch classification (the ‘rest’ class consisted of LGD and HGD).

In Chapter 6, by using only eight mouse clicks per image, we have achieved an AAUC value of 0.85 for the same task. In Chapter 7, with only binary labels indicating whether cancer was present, an AAUC of 0.81 was achieved.

8.2 Limitations

8.2.1 Polyp analysis in OPT

Compared to the emerging study of using OPT for colorectal polyp analysis, extensive literature exists on automatic analysis of more traditional histopathology images for a range of analysis tasks, clinical settings, and disease types. However, we are not aware of a suitable study with which to make direct comparisons (i.e., patch discrimination between LGD, HGD and ICA in colorectal polyps). The proposed system achieved promising performance in discriminating OPT image patches. Further improvements and incorporation into software tools would be needed to enable translation or adoption for clinical research. Larger datasets of annotated images would be needed to train and validate such tools because the visual appearance of dysplastic change is complex and the variations across polyps are large.

It is worth noting that inter-observer variation exists in polyp diagnosis with OPT [27]. In order to minimise the effect of such variation and uncertainty in our study, the ground truth was annotated within high confidence regions by only one experienced pathologist rather than trying to delineate accurate region boundaries. The annotated regions were meanwhile cross-checked and calibrated with the corresponding H&E slices to ensure the confidence of the obtained ground truth. Nevertheless, there may still exist some uncertainty in the ground truth. Quantifying this using multiple pathologists would be interesting for a future study.

The image regions outside the annotated areas may contain a mixture of dysplasia, invasive cancer, and other components (e.g., stroma and connective tissue). However,

the actual labels were unknown in the current dataset. This has prevented us from quantitatively validating our algorithms by labelling every voxel.

8.2.2 Partial annotations

Chapter 6 investigated training OPT patch classifiers using click annotations in a binary classification setting. In the training stage, each click and its context — either in terms of patch locations or in the patch feature space — were modelled individually without considering their mutual relations. In order to fully utilise the click information, this could be extended to incorporate multiple clicks per image by constructing a joint confidence map (e.g., using a mixture of Gaussians model).

8.2.3 Weakly supervised image analysis

In the multiple instance boosting algorithm proposed in Chapter 7, we have treated each cropped sub-volume of the OPT images as a bag. This was computationally feasible as each bag only contains about 200 instances. However, we were not able to use each OPT image as a bag because it would lead to a large number of instances per bag. In future work, it would be interesting to extend the algorithm to learn from very large bags.

The level of complexity in manual annotations was largely reduced, from delineating regions to mouse clicks, and further to image-level binary labels. However systematically quantifying the time taken to annotate OPT images, according to annotators' experiences and the complexity of polyps' appearances, can be an interesting future direction.

8.3 Future work

It would be important to compare and contrast the automated analysis of OPT with the gold standard H&E section using fully annotated datasets. Ideally, to make a rigorous

comparison of the microscopy imaging and analysis methods, detailed annotations would be needed for the images of the same polyp obtained with both imaging methods. Automatic histology analysis methods could be applied to generate segmentation scores with respect to each pixel or voxel for each image. The differences between the segmentation results and the annotations could be quantified and compared. Furthermore, the comparative studies of OPT images and H&E sections could include (1) polyp image registration using the H&E section and the virtual section of OPT, (2) transfer of the domain knowledge from H&E sections to analysis of OPT images, and vice versa, (3) fusion of histological analysis of both modalities to achieve a better diagnosis. For colorectal cancer research, these studies would potentially improve our understanding of colorectal polyps; for clinical applications, it would be interesting to investigate whether providing a synchronised visualisation of colorectal polyps using both H&E section and 3-D OPT images could improve the accuracy of diagnosis. OPT enables access to polyps' surface morphology and internal structure at the same time. In the future, one interesting direction is to combine morphological analysis of colorectal polyp surfaces (e.g., [156, 157]) with 3-D texture analysis.

The contextual ranking model was trained with a stochastic gradient descent method that can be updated online. Extending the algorithm to learn the model in an interactive manner would be an interesting direction. For example, the cancer detector could be trained in an active learning setting. The machine could identify a few hard or informative training examples and asks the annotator to provide clicks. These clicks can then be used to update the classification model online.

To reduce the annotation effort while training a good quality classification model, weakly supervised learning has shown promising results in this research. However, in both weakly supervised settings cubic image patches were used. In future, other approaches for generating regions or segments of interest using appearance information may be considered [142].

The feature extraction process is unsupervised in this research. In the future, learning feature extractors and classification models simultaneously would be an interesting direction. Recently deep learning techniques have yielded promising results in learning image representations from large datasets. Applying such techniques to OPT datasets would be an interesting direction. Moreover, instead of only learning from the three discrete class labels (LGD, HGD and ICA), the system output space could be extended to consider continuous labels reflecting the dysplastic changes. Further, the output could be extended to be a detailed clinical histology report with various items (classification with structured output).

The features extracted by the systems are difficult to interpret. One interesting research direction is to develop data mining algorithms that learn semantically meaningful visual clues or concepts from the images annotated at image-level [83, 106]. Since the accuracy of interpreting OPT images depends on the experience of the pathologist, visual concepts summarised from large-scale OPT datasets are potentially useful for pathologist training purposes. In order to train the pathologists to identify cancerous regions in OPT, a collection of representative regions searched from a large polyp database would be useful. Manually collecting the regions can be very laborious and difficult due to the large variations across the polyps. Automatically mining for the discriminative and semantically meaningful features or regions could be an efficient alternative.

Appendix A

List of publications

The following publications have resulted from work described in this thesis.

- *Classification of colorectal polyp regions in optical projection tomography*
  Li W, Zhang J, McKenna S J, Coats M, Carey F A (2013)
  International Symposium on Biomedical Imaging (ISBI), San Francisco

- *Learning from partially annotated OPT images by contextual relevance ranking*
  Li W, Zhang J, Zheng W, Coats M, Carey F A, McKenna S J (2013)
  International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Nagoya

- *Comparative analysis of feature extraction methods for colorectal polyp images in optical projection tomography*
  Li W, Coats M, Zhang J, McKenna S J (2013)
  Medical Image Understanding and Analysis (MIUA), Birmingham
  Best student paper

- *Multiple instance cancer detection by boosting regularised trees*
  Li W, Zhang J, McKenna S J (2015)
  International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich

- *Discriminating dysplasia: Optical tomographic texture analysis of colorectal polyps*
  Li W, Coats M, Zhang J, McKenna S J (2015)
  Medical Image Analysis 26 (2015) 57-59

The following is a list of contributions to other publications during the PhD period.

- *Multi-scale analysis of the surface morphology of colorectal polyps from optical tomography*
  Zhang J, Zhang J, Coats M, **Li W**, Carey F A, McKenna S J (2015)
  Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

- *HEp-2 cell classification using multi-resolution local patterns and ensemble SVMs*
  Manivannan S, **Li W**, Akbar S, Wang R, Zhang J, McKenna S J (2014)
  1st Workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A) International Conference on Pattern Recognition (ICPR), Stockholm
  Winner of International Contest on Performance Evaluation of I3A Systems, Task 1

- *HEp-2 specimen classification using multi-resolution local patterns and SVM*
  Manivannan S, **Li W**, Akbar S, Wang R, Zhang J, McKenna S J (2014)
  1st Workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A) International Conference on Pattern Recognition (ICPR), Stockholm
  Winner of International Contest on Performance Evaluation of I3A Systems, Task 2

- *Brain tumor region segmentation using local co-occurrence features and conditional random fields*
  Manivannan S, Shen H, **Li W**, Annunziata R, Hamad H, Wang R, Zhang J (2014)
  Brain Tumor Digital Pathology Segmentation Challenge MICCAI, Boston
  Best Performing Runner-up

- *An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens*
  Manivannan S, **Li W**, Akbar S, Wang R, Zhang J, McKenna S J (2015)
  Pattern Recognition

Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[2] N. Agarwal, Y. Xie, F. W. Patten, A. P. Reeves, and E. J. Seibel. DNA ploidy measure of feulgen-stained cancer cells using three-dimensional image cytometry. In *Healthcare Innovation Conference*, pages 6–9. IEEE, 2014.

[3] M. Y. Ahmad, A. Mohamed, Y. A. M. Yusof, and S. Md Ali. Colorectal cancer image classification using image pre-processing and multilayer perceptron. In *International Conference on Computer & Information Science (ICCIS)*, volume 1, pages 275–280. IEEE, 2012.

[4] A. Akselrod-Ballin, D. Bock, R. C. Reid, and S. K. Warfield. Accelerating image registration with the Johnson–Lindenstrauss lemma: Application to imaging 3-D neural ultrastructure with electron microscopy. *IEEE Transactions on Medical Imaging*, 30(7):1427–1438, 2011.

[5] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, 1:113–141, 2001.

[6] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, 57(3):665–674, 2010.

[7] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[8] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. A smart atlas for endomicroscopy using automated video retrieval. *Medical Image Analysis*, 15(4):460–476, 2011.

[9] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache. Learning semantic and visual similarity for endomicroscopy video retrieval. *IEEE Transactions on Medical Imaging*, 31(6):1276–1288, 2012.

[10] B. André, T. Vercauteren, A. Perchant, A. M. Buchner, M. B. Wallace, and N. Ayache. Endomicroscopic image retrieval and classification using invariant visual features. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 346–349. IEEE, 2009.

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 561–568. MIT, 2002.

[12] D. Arthur and S. Vassilvitskii. *k*-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. ACM, 2007.

[13] V. Atlamazoglou, D. Yova, N. Kavantzas, and S. Loukas. Texture analysis of fluorescence microscopic images of colonic tissue sections. *Medical and Biological Engineering and Computing*, 39(2):145–151, 2001.

[14] J. Ayoub, B. Granado, Y. Mhanna, and O. Romain. SVM based colon polyps classifier in a wireless active stereo endoscope. In *International Conference of Engineering in Medicine and Biology Society*, pages 5585–5588. IEEE, 2010.

[15] J. Ayoub, B. Granado, O. Romain, and Y. Mhanna. 3-D object recognition based on SVM and stereo-vision: application in endoscopic imaging. In *International Conference of Soft Computing and Pattern Recognition*, pages 198–201. IEEE, 2010.

[16] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *International Conference on Knowledge Discovery and Data Mining*, pages 245–250. ACM, 2001.

[17] C. Boutsidis, A. Zouzias, and P. Drineas. Random projections for *k*-means clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 298–306. MIT, 2010.

[18] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[19] H. Brenner, M. Kloor, and C. P. Pox. Colorectal cancer. *The Lancet*, 383(9927):1490 – 1502, 2014.

[20] T. Brosch and R. Tam. Manifold learning of brain MRIs by deep learning. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 633–640. Springer, 2013.

[21] Bruker microCT. "Bioptonics" optical projection tomography. http://www.skyscan.be/products/OEM.htm, 2005. Accessed on 08/04/2015.

[22] Cancer Research UK. Bowel cancer key stats. http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/bowel-cancer/cancerstats-key-facts-on-bowel-cancer, 2014. Accessed on 10/04/2015.

[23] A. Chaddad, C. Tanougast, A. Dandache, and A. Bouridane. Extraction of Haralick features from segmented texture multispectral bio-images for detection of colon cancer cells. In *International Conference on Informatics and Computational Intelligence*, pages 55–59. IEEE, 2011.

[24] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010.

[25] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[26] Y. Chen, C.-P. Liang, Y. Liu, A. Fischer, A. Parwani, and L. Pantanowitz. Review of advanced imaging techniques. *Journal of Pathology Informatics*, 3(1):22, 2012.

[27] M. Coats, S. Wedden, R. Keogh, R. Steele, and F. Carey. Diagnosis of colorectal polyps using optical projection tomography – how well do pathologists agree? *United European Gastroenterology Journal*, 1:(Supplement 1) A233, 2013.

[28] M. V. Coats, S. E. Wedden, J. Farrell, G. Cranston, L. Mitchell, J. Wilson, R. J. Steele, and F. A. Carey. Optical projection tomography: Can it help diagnose the colorectal polypoid cancer? *Gastroenterology*, 142(5):S178–S179, 2012.

[29] A. Cohen, E. Rivlin, I. Shimshoni, and E. Sabo. Colon biopsy classification using crypt architecture. In *Machine Learning in Medical Imaging*, pages 182–189. Springer, 2014.

[30] A. Cohen, E. Rivlin, I. Shimshoni, and E. Sabo. Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation. *Computerized Medical Imaging and Graphics*, 43(0):150 – 164, 2015.

[31] P. M. Colucci, S. H. Yale, and C. J. Rall. Colorectal polyps. *Clinical medicine & research*, 1(3):261–262, 2003.

[32] D. Cunningham, W. Atkin, H.-J. Lenz, H. T. Lynch, B. Minsky, B. Nordlinger, and N. Starling. Colorectal cancer. *The Lancet*, 375(9719):1030 – 1047, 2010.

[33] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[34] I. Dawson. Histological typing of intestinal tumours. International histological classification of tumours. *Journal of Clinical Pathology*, 30(7):685, 1977.

[35] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 40:837–845, 1988.

[36] H. Deng and G. Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013.

[37] A. Depeursinge, A. Foncubierta-Rodriguez, D. Van De Ville, and H. Müller. Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities. *Medical Image Analysis*, 18(1):176–196, 2014.

[38] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212, 1996.

[39] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[40] S. Doyle, J. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinformatics*, 12(1):1–14, 2011.

[41] M. Dundar, S. Badve, V. Raykar, R. Jain, O. Sertel, and M. Gurcan. A multiple instance learning approach toward optimal classification of pathology slides. In *International Conference on Pattern Recognition (ICPR)*, pages 2732–2735. IEEE, 2010.

[42] A. N. Esgiar, R. Naguib, M. K. Bennett, and A. Murray. Automated feature extraction and identification of colon carcinoma. *Analytical and Quantitative Cytology and Histology*, 20(4):297–301, 1998.

[43] A. N. Esgiar, R. N. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):54–58, 2002.

[44] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[45] J. Fehr and H. Burkhardt. 3D rotation invariant local binary patterns. In *International Conference on Pattern Recognition (ICPR)*, pages 1–4. IEEE, 2008.

[46] C. Fenger, M. Bak, O. Kronborg, and H. Svanholm. Observer reproducibility in grading dysplasia in colorectal adenomas: comparison between two different grading systems. *Journal of Clinical Pathology*, 43(4):320–324, 1990.

[47] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray. Globocan 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase no. 11 [internet]. Lyon, France: International agency for research on cancer. Available from: http://globocan.iarc.fr, accessed on 04/march/2014., 2013.

[48] B. P. Flannery, H. W. Deckman, W. G. Roberge, and K. L. D'AMICO. Three-dimensional X-ray microtomography. *Science*, 237(4821):1439–1444, 1987.

[49] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[50] J. J. Fu, Y.-W. Yu, H.-M. Lin, J.-W. Chai, and C. C.-C. Chen. Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging. *Computerized Medical Imaging and Graphics*, 38(4):267–275, 2014.

[51] Z. Fu, A. Robles-Kelly, and J. Zhou. Milis: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):958–977, 2011.

[52] L. Gan Lim, R. N. Naguib, E. P. Dadios, and J. Avila. Implementation of GA-KSOM and ANFIS in the classification of colonic histopathological images. In *TENCON 2012-2012 IEEE Region 10 Conference*, pages 1–5. IEEE, 2012.

[53] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *Biomedical Engineering, IEEE Reviews in*, 2:147–171, 2009.

[54] M. Häfner, L. Brunauer, H. Payer, R. Resch, A. Gangl, A. Uhl, F. Wrba, and A. Vécsei. Computer-aided classification of zoom-endoscopical images using Fourier filters. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):958–970, 2010.

[55] M. Häfner, A. Gangl, R. Kwitt, A. Uhl, A. Vécsei, and F. Wrba. Improving pit–pattern classification of endoscopy images by a combination of experts. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 247–254. Springer, 2009.

[56] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Endoscopic image classification using edge-based features. In *International Conference on Pattern Recognition (ICPR)*, pages 2724–2727. IEEE, 2010.

[57] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, and F. Wrba. Pit pattern classification using extended local binary patterns. In *International Conference on Information Technology and Applications in Biomedicine*, pages 1–4. IEEE, 2009.

[58] M. Häfner, A. Gangl, F. Wrba, K. Thonhauser, H.-P. Schmidt, C. Kastinger, and A. Uhl. Comparison of *k*-NN, SVM, and NN in pit pattern classification of zoom-endoscopic colon images using co-occurrence histograms. In *International Symposium on Image and Signal Processing and Analysis*, pages 516–521. IEEE, 2007.

[59] M. Hafner, C. Kendlbacher, W. Mann, W. Taferl, F. Wrba, A. Gangl, A. Vecsei, and A. Uhl. Pit pattern classification of zoom-endoscopic colon images using histogram techniques. In *Nordic Signal Processing Symposium (NORSIG)*, 2006.

[60] M. Häfner, R. Kwitt, A. Uhl, F. Wrba, A. Gangl, and A. Vécsei. Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps. *Pattern Recognition*, 42(6):1180–1191, 2009.

[61] M. Häfner, M. Liedlgruber, S. Maimone, A. Uhl, and F. Wrba. Evaluation of cross-validation protocols for the classification of endoscopic images of colonic polyps. In *International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, 2012.

[62] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Color treatment in endoscopic image classification using multi-scale local color vector patterns. *Medical Image Analysis*, 16(1):75–86, 2012.

[63] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Delaunay triangulation-based pit density estimation for the classification of polyps in high-magnification chromo-colonoscopy. *Computer Methods and Programs in Biomedicine*, 107(3):565–581, 2012.

[64] M. Häfner, A. Uhl, A. Vécsei, G. Wimmer, and F. Wrba. Complex wavelet transform variants and discrete cosine transform for scale invariance in magnification-endoscopy image classification. In *International Conference on Information Technology and Applications in Biomedicine*, pages 1–5. IEEE, 2010.

[65] P. W. Hamilton, P. H. Bartels, D. Thompson, N. H. Anderson, R. Montironi, and J. M. Sloan. Automated location of dysplastic fields in colorectal histology using image texture analysis. *The Journal of Pathology*, 182(1):68–75, 1997.

[66] S. Hamilton and L. Aaltonen. *Pathology and genetics of tumours of the digestive system*. IARC press Lyon, 2000.

[67] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*. Springer, 2009.

[68] X. He and E. C. Frey. The meaning and use of the volume under a three-class ROC surface (VUS). *IEEE Transactions on Medical Imaging*, 27(5):577–588, 2008.

[69] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 115–132. MIT, 1999.

[70] P. Hermanek. Dysplasia in the gastrointestinal tract: definition and clinical significance. *Surgical Endoscopy*, 1(1):5–10, 1987.

[71] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, and C. A. Puliafito. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991.

[72] A. Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1984), 2012.

[73] A. Hyvarinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[74] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural image statistics*, volume 39. Springer, 2009.

[75] Y. Imai, S. Kudo, O. Tsuruta, T. Fujii, S. Hayashi, S. Tanaka, and T. Terai. Problems and clinical significance of v type pit pattern diagnosis: report on round-table consensus meeting. *Early Colorectal Cancer*, 5:595–613, 2001.

[76] E. Jurrus, A. R. Paiva, S. Watanabe, J. R. Anderson, B. W. Jones, R. T. Whitaker, E. M. Jorgensen, R. E. Marc, and T. Tasdizen. Detection of neuron membranes in electron microscopy images using a serial neural network architecture. *Medical Image Analysis*, 14(6):770–783, 2010.

[77] H. Kalkan, M. Nap, R. Duin, and M. Loog. Automated colorectal cancer diagnosis for whole-slice histopathology. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 550–557. Springer, 2012.

[78] H. Kalkan, M. Nap, R. P. Duin, and M. Loog. Automated classification of local patches in colon histopathology. In *International Conference on Pattern Recognition (ICPR)*, pages 61–64. IEEE, 2012.

[79] M. Kandemir, A. Feuchtinger, A. Walch, and F. A. Hamprecht. Digital pathology: Multiple instance learning can detect Barrett's cancer. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1348–1351. IEEE, 2014.

[80] M. Kandemir, C. Zhang, and F. A. Hamprecht. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 228–235. Springer, 2014.

[81] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu. Colorectal tumours and pit pattern. *Journal of Clinical Pathology*, 47(10):880–885, 1994.

[82] R. Kwitt and A. Uhl. Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In *International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[83] R. Kwitt, N. Vasconcelos, N. Rasiwasia, A. Uhl, B. Davis, M. Häfner, and F. Wrba. Endoscopic image analysis in semantic space. *Medical Image Analysis*, 16(7):1415–1422, 2012.

[84] Q. V. Le, J. Han, J. W. Gray, P. T. Spellman, A. Borowsky, and B. Parvin. Learning invariant features of tumor signatures. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 302–305. IEEE, 2012.

[85] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3361–3368. IEEE, 2011.

[86] A. Leeper, J. Farrell, J. Dixon, S. Wedden, D. Harrison, and E. Katz. Long-term culture of human breast cancer specimens and their analysis using optical projection tomography. *Journal of Visualized Experiments*, (53):193–208, 2010.

[87] P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *International Conference on Knowledge Discovery and Data Mining*, pages 287–296. ACM, 2006.

[88] W. Li, M. Coats, J. Zhang, and S. J. McKenna. Comparative analysis of feature extraction methods for colorectal polyp images in optical projection tomography. In *Medical Image Understanding and Analysis (MIUA)*, pages 67–72, 2013.

[89] W. Li, J. Zhang, and S. J. McKenna. Multiple instance cancer detection by boosting regularsed trees. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, page 0. in press, 2015.

[90] W. Li, J. Zhang, S. J. McKenna, M. Coats, and F. A. Carey. Classification of colorectal polyp regions in optical projection tomography. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 736–739. IEEE, 2013.

[91] W. Li, J. Zhang, W.-S. Zheng, M. Coats, F. A. Carey, and S. J. McKenna. Learning from partially annotated OPT images by contextual relevance ranking. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 429–436. Springer, 2013.

[92] S. Liao, Y. Gao, A. Oto, and D. Shen. Representation learning: A unified deep learning framework for automatic prostate MR segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 254–261. Springer, 2013.

[93] M. Liedlgruber and A. Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *IEEE Reviews in Biomedical Engineering*, 4:73–88, 2011.

[94] L. Liu and P. Fieguth. Texture classification from random features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):574–586, 2012.

[95] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, and J. M. Rehg. Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. *Medical Image Analysis*, 15(5):748–759, 2011.

[96] M. B. Loughrey and N. A. Shepherd. The pathology of bowel cancer screening. *Histopathology*, 66(1):66–77, 2015.

[97] J. Luo and C. Xiong. DiagTest3Grp: An R package for analyzing diagnostic tests with three ordinal groups. *Journal of Statistical Software*, 51(3):1, 2012.

[98] S. Maji, A. C. Berg, and J. Malik. Efficient classification for additive kernel SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):66–77, 2013.

[99] T. Majtner and D. Svoboda. Comparison of 3D texture-based image descriptors in fluorescence microscopy. In *Combinatorial Image Analysis*, pages 186–195. Springer, 2014.

[100] D. Majumdar, J. Patnick, C. Nickerson, and M. Rutter. Analysis of colorectal polyps detected in the English NHS bowel cancer screening programme with emphasis on advanced adenoma and polyp cancer detected. *Gut*, 61(Suppl 2):A67–A67, 2012.

[101] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna. An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens. *Pattern Recognition*, 2015.

[102] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 570–576. MIT, 1998.

[103] K. Masood and N. Rajpoot. Spatial analysis for colon biopsy classification from hyperspectral imagery. *The Annals of the BMVA*, 2008(4):1–16, 2008.

[104] K. Masood and N. Rajpoot. Texture based classification of hyperspectral colon biopsy samples using CLBP. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1011–1014. IEEE, 2009.

[105] M. McCann, J. Ozolek, C. Castro, B. Parvin, and J. Kovacevic. Automated histology analysis: Opportunities for signal processing. *Signal Processing Magazine, IEEE*, 32(1):78–87, Jan 2015.

[106] M. T. McCann, R. Bhagavatula, M. C. Fickus, J. A. Ozolek, and J. Kovacevic. Automated colitis detection from endoscopic biopsies as a tissue screening tool in diagnostic pathology. In *IEEE International Conference on Image Processing (ICIP)*, pages 2809–2812. IEEE, 2012.

[107] D. Mitrea, S. Nedevschi, and R. Badea. The role of the textural microstructure cooccurrence matrices in the classification of the abdominal tumors, based on ultrasound images. In *International Conference on Intelligent Computer Communication and Processing*, pages 187–190. IEEE, 2014.

[108] Y. Muller, S. Gupta, P. Morel, S. Borot, F. Bettens, M. Truchetet, J. Villard, J. Seebach, D. Holmberg, and C. Toso. Transplanted human pancreatic islets after long-term insulin independence. *American Journal of Transplantation*, 13(4):1093–1097, 2013.

[109] T. Muto, H. Bussey, and B. Morson. Pseudo-carcinomatous invasion in adenomatous polyps of the colon and rectum. *Journal of Clinical Pathology*, 26(1):25, 1973.

[110] NHS BCSP. Reporting lesions in the NHS bowel cancer screening programme - guidelines from the bowel cancer screening programme pathology group. http://www.cancerscreening.nhs.uk/bowel/publications/nhsbcsp01.pdf, 2007. Accessed on 13/07/2014.

[111] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[112] G. Olgun, C. Sokmensuer, and C. Gunduz-Demir. Local object patterns for the representation and classification of colon tissue images. *IEEE Journal of Biomedical and Health Informatics*, 18(4):1390–1396, 2014.

[113] K. Oppedal, K. Engan, D. Aarsland, M. Beyer, O. B. Tysnes, and T. Eftestol. Using local binary pattern to classify dementia in MRI. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 594–597. IEEE, 2012.

[114] E. Ozdemir and C. Gunduz-Demir. A hybrid classification model for digital pathology using structural and statistical pattern recognition. *IEEE Transactions on Medical Imaging*, 32(2):474–483, 2013.

[115] E. Ozdemir, C. Sokmensuer, and C. Gunduz-Demir. A resampling-based markovian model for automated colon cancer diagnosis. *IEEE Transactions on Biomedical Engineering*, 59(1):281–289, 2012.

[116] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[117] K. Rajpoot and N. Rajpoot. SVM optimization for hyperspectral colon tissue cell classification. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 829–837. Springer, 2004.

[118] S. Rathore, M. Hussain, A. Ali, and A. Khan. A recent survey on colon cancer detection techniques. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3):545–563, 2013.

[119] S. Rathore, M. Hussain, M. A. Iftikhar, and A. Jalil. Ensemble classification of colon biopsy images based on information rich hybrid features. *Computers in Biology and Medicine*, 47:76–92, 2014.

[120] S. Rathore, M. Hussain, and A. Khan. Automated colon cancer detection using hybrid of novel geometric features and some traditional features. *Computers in Biology and Medicine*, (0):0, 2015. in press.

[121] S. Rathore, M. A. Iftikhar, M. Hussain, and A. Jalil. Classification of colon biopsy images based on novel structural features. In *International Conference on Emerging Technologies*, pages 1–6. IEEE, 2013.

[122] J. D. Rennie and N. Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *IJCAI multidisciplinary workshop on advances in preference handling*, pages 180–186, 2005.

[123] E. Rodriguez-Diaz, D. A. Castanon, S. K. Singh, and I. J. Bigio. Spectral classifier design with ensemble classifiers and misclassification-rejection: application to elastic-scattering spectroscopy for detection of colonic neoplasia. *Journal of Biomedical Optics*, 16(6):067009–067009–16, 2011.

[124] N. L. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680. MIT, 2012.

[125] X. Shao, W. Zheng, and Z. Huang. Near-infrared autofluorescence spectroscopy for in vivo identification of hyperplastic and adenomatous polyps in the colon. *Biosensors and Bioelectronics*, 30(1):118–122, 2011.

[126] R. Sharma. Microimaging of hairless rat skin by magnetic resonance at 900 MHz. *Magnetic Resonance Imaging*, 27(2):240–255, 2009.

[127] J. Sharpe. Optical projection tomography. *Annual Review of Biomedical Engineering*, 6:209–228, 2004.

[128] J. Sharpe. Optical projection tomography. In *Advanced Imaging in Biology and Medicine*, pages 199–224. Springer, 2009.

[129] J. Sharpe, U. Ahlgren, P. Perry, B. Hill, A. Ross, J. Hecksher-Sørensen, R. Baldock, and D. Davidson. Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science*, 296(5567):541–545, 2002.

[130] J. Shuttleworth, A. Todman, R. Naguib, B. Newman, and M. Bennett. Multiresolution colour texture analysis for classifying colon cancer images. In *Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, Proceedings of the Second Joint*, volume 2, pages 1118–1119, 2002.

[131] R. Siegel, C. DeSantis, K. Virgo, K. Stein, A. Mariotto, T. Smith, D. Cooper, T. Gansler, C. Lerro, and S. Fedewa. Cancer treatment and survivorship statistics. *Ca: A Cancer Journal for Clinicians*, 62(4):220–241, 2012.

[132] A. C. Simsek, A. B. Tosun, C. Aykanat, C. Sokmensuer, and C. Gunduz-Demir. Multilevel segmentation of histopathological images using cooccurrence of tissue objects. *IEEE Transactions on Biomedical Engineering*, 59(6):1681–1690, 2012.

[133] B. Song, G. Zhang, H. Lu, H. Wang, W. Zhu, P. J. Pickhardt, and Z. Liang. Volumetric texture features from higher-order images for diagnosis of colon lesions via CT colonography. *International Journal of Computer Assisted Radiology and Surgery*, 9(6):1021–1031, 2014.

[134] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning (ICML)*, volume 32, pages 1611–1619, 2014.

[135] Y. Song, D. Treanor, A. Bulpitt, and D. Magee. 3D reconstruction of multiple stained histology images. *Journal of Pathology Informatics*, 4(2):7, 2013.

[136] L. Sorensen, S. B. Shaker, and M. De Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Transactions on Medical Imaging*, 29(2):559–569, 2010.

[137] T. Stehle, R. Auer, S. Gross, A. Behrens, J. Wulff, T. Aach, R. Winograd, C. Trautwein, and J. Tischendorf. Classification of colon polyps in nbi endoscopy using vascularization features. volume 7260, pages 72602S–72602S–12, 2009.

[138] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyaki, and S. Tanaka. Computer-aided colorectal tumor classification in NBI endoscopy using local features. *Medical Image Analysis*, 17(1):78–100, 2013.

[139] J. Tischendorf, S. Gross, R. Winograd, H. Hecker, R. Auer, A. Behrens, C. Trautwein, T. Aach, and T. Stehle. Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study. *Endoscopy*, 42(3):203–207, 2010.

[140] A. B. Tosun and C. Gunduz-Demir. Graph run-length matrices for histopathological image segmentation. *IEEE Transactions on Medical Imaging*, 30(3):721–732, 2011.

[141] A. B. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz-Demir. Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection. *Pattern Recognition*, 42(6):1104 – 1112, 2009.

[142] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[143] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–691–8. IEEE, 2003.

[144] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[145] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

[146] B. C. Wallace and I. J. Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and Information Systems*, 41(1):33–52, 2014.

[147] T. Wilson. Confocal microscopy. *Academic Press: London, etc*, 426:1–64, 1990.

[148] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1626–1630. IEEE, 2014.

[149] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 964–971. IEEE, 2012.

[150] Y. Xu, J.-Y. Zhu, I. Eric, C. Chang, M. Lai, and Z. Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014.

[151] H. Yoshida and J. Näppi. CAD in CT colonography without and with oral contrast agents: progress and challenges. *Computerized Medical Imaging and Graphics*, 31(4):267–284, 2007.

[152] H. Yoshida, J. Näppi, P. MacEneaney, D. T. Rubin, and A. H. Dachman. Computer-aided diagnosis scheme for detection of polyps at CT colonography. *Radiographics*, 22(4):963–979, 2002.

[153] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31:1116–1128, 2006.

[154] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1417–1424. MIT, 2005.

[155] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[156] J. Zhang, S. J. McKenna, J. Zhang, M. Coats, and F. Carey. Analysing the surface morphology of colorectal polyps: Differential geometry and pit pattern prediction. In *Medical Image Understanding and Analysis (MIUA)*, pages 67–72, 2014.

[157] J. Zhang, J. Zhang, M. Coats, W. Li, F. Carey, and S. McKenna. Multi-scale analysis of the surface morphology of colorectal polyps from optical tomography. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 0(0):1–11, 2015.

[158] Q. Zhang and S. A. Goldman. EM-DD: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1073–1080. MIT, 2001.

[159] D. Zhao, Y. Chen, and N. Correa. Automated classification of human histological images, a multiple-instance learning approach. In *Life Science Systems and Applications Workshop, IEEE/NLM*, pages 1–2. IEEE, 2006.

[160] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.